ABSTRACT

OSMAN, MUHTARJAN. Bayesian Noninferiority Testing and Nonparametric Survival Regression using Bernstein Polynomials. (Under the direction of Dr. Sujit K. Ghosh.)

In this dissertation we study two distinctly important topics in biostatistics. In both, Bernstein polynomials are used as building blocks for the statistical models. The first part involves the development of a new Bayesian procedure for hypothesis testing of noninferiority. A semiparametric testing approach based on Bayes factor is developed for non-inferiority trials with binary endpoints. The proposed method is shown to work for a broad class of hypotheses by accommodating a variety of dissimilarity measures between two binomial parameters. Two of the unique features of the proposed testing procedure include: (i) construction of a flexible class of conjugate priors using a mixture of Beta densities to maintain approximate equality of prior probabilities of the competing hypotheses; and (ii) automatic determination of the cut-off value of the Bayes factor to facilitate the decision making process. In contrast to the use of Jeffreys' rule of thumb, two forms of total weighted average error criteria are used to determine the cut-off value. The second part of the dissertation focuses on nonparametric regression models for right-censored data. We present a new nonparametric regression model for the conditional hazard rate using a suitable sieve of Bernstein polynomials. The proposed nonparametric methodology has three key features: (i) the smooth estimator of the conditional hazard rate is shown to be a unique solution of a strictly convex optimization problem for a wide range of applications; making it computationally attractive, (ii) the model is shown to nest the popular Cox proportional hazard model, and (iii) large sample properties including consistency and convergence rates are established under a set of mild regularity conditions. Finally, we proposed a Bayesian nonparametric model for the conditional hazard function based on Bernstein polynomials with varying degrees. The most important feature of the proposed Bayesian method is its capability of modeling the uncertainty about the order of Bernstein polynomials using reversible jump Markov Chain Monte Carlo (RJ-MCMC). This has been a

great challenge for the proposed method from a frequentist perspective because of the lack of cross-validation procedures for censored data in regression settings.

Bayesian Noninferiority Testing and Nonparametric Survival Regression using Bernstein Polynomials

by
Muhtarjan Osman

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2011

APPROVED BY:

Dr. Brian Reich	Dr. Howard Bondell
Dr. Subhashis Ghoshal	Dr. Sujit K. Ghosh Chair of Advisory Committee

DEDICATION

To my parents, my wife, and my baby daughter

BIOGRAPHY

Muhtarjan Osman graduated from high school in 1999. He received his Bachelor's degree from Peking University in 2003. Then he went to the University of California at Davis and got his Master's degree in 2006. In August 2006 he joined the Department of Statistics at North Carolina State University for the doctoral level study in statistics.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude and appreciation to my advisor Dr. Sujit K. Ghosh for his guidance, support, and encouragement. I feel so fortunate to have such a great mentor. I want to thank my advisory committee members Dr. Howard Bondell, Dr. Subhashis Ghoshal, and Dr. Brian Reich for their insightful comments and helpful suggestions. Also, I want to thank Dr. Bruce Branson from the College of Management for serving in my dissertation committee.

I am grateful to the departmental support in terms of graduate funding and state of art research facilities. It is my honor to study in such a great department and to receive the finest graduate education from world-class scholars in statistics. The department has one of the most well-organized graduate programs. Here I want to thank the graduate program directors Dr. Pam Arroway, Dr. Jacqueline Hughes-Oliver, and Dr. Sujit K. Ghosh for their help and advice. I also want to thank all administrative staff in the department, particularly Mr. Adrian Blue for his extraordinary work helping graduate students.

Finally, I owe a great debt to my parents and my wife for their love and support throughout my graduate study. Without them I would not be able to reach the point that I am right now.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Semiparametric Bayesian Testing Procedure for Noninferiority Trials with Binary Endpoints

1.1 Introduction

In some medical applications, an active control is used instead of placebo to compare against the experimental treatment due to either ethical or feasibility concerns. Since the efficacy is not the only measure of the medical and commercial value of a treatment, the experimental treatment will be considered beneficial if its efficacy is not inferior to the active control by some pre-specified standard given that the experimental treatment is believed to have other advantages such less toxicity, easier administration, or lower cost.

In clinical trials with two-sample binomial data, the null and alternative hypotheses for noninferiority testing can often be written as $H_0 : \theta_2 - \theta_1 \leq -\delta$ vs. $H_1 : \theta_2 - \theta_1 > -\delta$, where θ_1 and θ_2 denote the success rates of the active control group and the experimental group, respectively, and $\delta > 0$ is a pre-specified noninferiority margin. Based on the asymptotic normality at the boundary of the hypotheses, a one-sided testing procedure is used to evaluate the evidence

against the inferiority of the experimental treatment (Blackwelder, 1982). This approach is further refined by Farrinton and Manning (1990) by replacing the sample proportions in the standard error with the restricted maximum likelihood estimates of θ_1 and θ_2 on the boundary of the hypotheses. Since both θ_1 and θ_2 belong to the unit interval $[0, 1]$, alternative dissimilarity measures such as relative risk (θ_2/θ_1) and odds ratio ($\frac{\theta_2(1-\theta_1)}{(1-\theta_2)\theta_1}$) can also be used to test noninferiority. Asymptotic testing procedure for hypotheses using relative risk and odds ratio can be constructed using the standard "delta" method approximation (see Tu, 1998; Laster and Johnson, 2006), and the corresponding one-sided testing procedures based on asymptotic normality will follow.

For small sample data, some exact procedures have been proposed (see e.g., Chan, 2003), however, Röhmel (2005) pointed out that such methods "are not yet mature for general use" due to the violation of coherence of the exact p-value. No matter what dissimilarity measures are used or whether the test procedure is based on asymptotic or exact distributions, the frequentist procedures often use p-value as the measure of evidence against the null hypothesis. In the context of noninferiority testing, p-value obtained by the above procedures is simply the probability of observing more extreme data than the test statistic under the distribution obtained on the boundary of the hypotheses. As summarized in Ghosh et al. (2006, chap. 6), there are several drawbacks of using p-value as the measure of evidence. First, the more extreme observation is not a part of data. Second, the probability is only evaluated on the boundary not on the entire space that satisfies the null hypothesis. Even though this can be justified in some situations by using Barnard's convexity argument (see Röhmel, 2005), the rationality is not yet clear in some other cases such as tests based on relative risk or odds ratio. Finally, p-value does not take into account the likelihood of the data under the alternative hypothesis.

For the sample size calculation in noninferiority tests, the frequentist procedure is to find a sample size such that the probability of the rejection region under the alternative hypothesis reaches some pre-specified level such as 0.8. Since in the frequentist framework one can not evaluate the probability over the entire region of the alternative hypothesis, usually a line (e.g.,

$\theta_2 = \theta_1$), which is only a small portion of the region of the alternative hypothesis, is used when calculating the statistical power. In other words, the frequentist procedure treats the composite null and alternative hypotheses as if they were simple null and alternative (e.g., $H_0^* : \theta_2 - \theta_1 = -\delta$ vs. $H_1^* : \theta_2 - \theta_1 = 0$). The problem becomes more severe when the likelihood (or some conditional likelihood) can not be expressed as a function of the dissimilarity measures (e.g., $\theta_1 - \theta_2$ or θ_1/θ_2 etc.).

Bayesian testing procedures are usually based on the posterior probability of the competing hypotheses or on the Bayes factor, which is defined as the ratio of posterior odds of the hypotheses to the corresponding prior odds. Compared to p-value, the posterior probability or Bayes factor provide more comprehensive measures of evidence against the null hypothesis because they are strictly based on the likelihood principle and takes into the uncertainty of both of the competing hypotheses. When prior probabilities of the null and alternative hypotheses are roughly equal, the test based on the posterior probability is roughly equivalent to the test based on the Bayes factor. Through out this paper, we use Bayes factor as the test statistic by using a flexible class of priors that achieve approximate equality of prior probabilities of the competing hypotheses. We define the Bayes factor in favor of the alternative hypothesis, so larger values of the Bayes factor are regarded as stronger evidence against the null hypothesis.

A prior probability for unknown parameters is essential and plays a crucial role in a Bayesian procedure especially in case of testing hypotheses. Usually, a parametric conjugate Beta prior is assigned to binomial parameters. For example, Wellek (2005) studied the test for noninferiority using posterior probability of the alternative hypothesis with Jeffreys prior for θ_1 and θ_2. Also, a test based on Bayes factor using parametric conjugate priors is studied by Williamson (2007) in the context of equivalence tests. The choice of the form of prior density as a Beta distribution is mainly due to the computational convenience brought by conjugacy. However, such choices are often criticized by many non-Bayesian practitioners. We develop a class of semiparametric conjugate priors using a suitable mixture of Beta densities following the work of Diaconis and Ylvisaker (1985). As a result, we relax the assumption on the prior of the binomial parameters to

approximate any continuous probability density function on the support of $[0, 1]$. Such priors are constructed in a flexible way such that they not only can incorporate historical data or expert opinion but also can maintain the objective of equal prior probabilities of the competing hypotheses when there is no prior information available.

In order to use the Bayes factor from a regulatory perspective, one needs determine a cut-off value of the Bayes factor that enables to make a decision in favour of one of the competing hypotheses. Usually, in the Bayesian framework the popular choice of the cut-off value is based on Jeffreys' empirical scale of evidence (see Kass and Raftery, 1995). Such empirical rules may be adequate in some cases, however, in most medical applications which are highly regulated, such empirical rules of thumb for the cut-off values are not widely used because they are not designed to control any type of decision error.

In this paper, we propose a semiparametric procedure for testing noninferiority to address the existing limitations of current approaches that we have mentioned above. The proposed procedure is a hybrid Bayesian/frequentist approach in the sense that Bayes factor is used as the measure of evidence against the null hypothesis while the frequentist ideas of controlling type I and II errors are extended to determine the cut-off value of Bayes factor to make a decision.

The idea of "borrowing" the concept of type I error to determine the cut-off value of Bayes factor in hypothesis testing has been considered by Weiss (1997). The Bayesian extension of type I error can be regarded as prior-weighted average of the frequentist type I error over the region of null hypothesis. Correspondingly, the sample size can be determined using the concept of Bayesian power function (see also Spiegelhalter and Freedman, 1986). In the literature this is often referred as the hybrid classical and Bayesian approach for sample size calculation. Compared to fully Bayesian methods such as the HPD-based approach (Joseph et al., 1997) and the decision-theoretic approach (Lindley, 1997), the hybrid approach has the advantage that it is easier to implement in applications especially in highly regulated fields where type I and II errors are the most focused criteria at the design stage. On the other hand, since the

Bayesian type I error is a weighted average of classical type I error, so it still can not serve the purpose of controlling the maximum error that is required by most regulatory agencies. Alternatively, in this paper we propose choosing the critical value of the Bayes factor that minimizes a total weighted error criterion. Two versions of such criterion are studied: one is based on controlling the average of usual frequentist type I and II errors as in Weiss (1997), the other is similar in spirit to that of Lee and Zelen (2000). The rest of the paper proceeds as follows. A detailed description of the proposed semiparametric Bayesian testing procedure is given in Section 1.2. In Section 2.4, the proposed method is demonstrated through simulated data as well as real data from a streptococcal pharyngitis trial. Finally, we conclude with some discussions in Section 1.4.

1.2 Semiparametric Bayesian Approach for Noninferiority Test

Considering a general framework, suppose we observe data $X \sim f(x|\theta)$ where $f(x|\theta)$ denotes the conditional density of the data (vector) X given the parameter $\theta \in \Theta$. Let $\theta \sim \pi(\cdot)$, where $\pi(\cdot)$ is a (prior) probability density defined on the parameter space Θ. Consider the problem of comparing two competing hypotheses:

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1,$$

where $\Theta_0 \cap \Theta_1 = \emptyset$; and $\Theta_0 \cup \Theta_1 \subseteq \Theta$. Assume that $Pr[\theta \in \Theta_j] = \int_{\Theta_j} \pi(\theta) d\theta > 0$ for $j = 0, 1$. Then the Bayes factor (BF) in favor of the alternative hypothesis is defined as the ratio of the posterior odds to the prior odds:

$$BF(X) = \frac{Pr[\theta \in \Theta_1|X]}{Pr[\theta \in \Theta_0|X]} \cdot \frac{Pr[\theta \in \Theta_0]}{Pr[\theta \in \Theta_1]}.$$

With the above definition of Bayes factor one would reject the null hypothesis if $BF(X) > B_0$ for some cut-off value $B_0 > 0$.

In the setting of a randomized clinical trial with two treatment arms, suppose X_1 out of n_1 patients who received the standard treatment had a "success" event, whereas the number of successes in the experimental treatment group of size n_2 is denoted by X_2. Assume that the two treatment success counts are independent, i.e., $X_i|\theta_i \overset{ind.}{\sim} Bin(n_i, \theta_i)$ for $i = 1, 2$, and θ_1 and θ_2 are the true success rates of the active control and the experimental treatment, respectively. Now we consider a general form of the null and alternative hypotheses for noninferiority tests as:

$$H_0 : \theta_2 \leq g(\theta_1, \rho) \text{ vs. } H_1 : \theta_2 > g(\theta_1, \rho), \tag{1.1}$$

where ρ is a pre-determined real-valued quantity and $g(\cdot, \cdot)$ is a continuous function of θ_1 and ρ. Often $g(\theta_1, \rho)$ is an increasing function of θ_1 for a given fixed value of ρ. When $g(\theta_1, \rho) = \theta_1 - \rho$, it leads to the null hypothesis of comparing the differences $H_0 : \theta_2 - \theta_1 \leq -\rho$, where $\rho > 0$ is usually called the noninferiority margin. Also $g(\theta_1, \rho) = \rho\theta_1$ leads to the null hypothesis to compare the relative risk $H_0 : \theta_2 \leq \rho\theta_1$, and $g(\theta_1, \rho) = \frac{\rho\theta_1}{1-\theta_1+\rho\theta_1}$ leads to the null hypothesis of comparing the odds ratio $H_0 : \frac{\theta_2}{1-\theta_2} \leq \rho\frac{\theta_1}{1-\theta_1}$ where ρ is often chosen to $\rho < 1$. The hypotheses (1.1) can be equivalently expressed as $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, where $\theta = (\theta_1, \theta_2)$, $\Theta_0 = \{(\theta_1, \theta_2) \in [0, 1]^2 : \theta_2 \leq g(\theta_1, \rho)\}$, and $\Theta_1 = \{(\theta_1, \theta_2) \in [0, 1]^2 : \theta_2 > g(\theta_1, \rho)\}$.

1.2.1 Prior Specification

Following the standard Bayesian inferential procedure, the binomial parameters are assigned independent prior distributions: $\theta_1 \sim \pi_1(\cdot)$ and $\theta_2 \sim \pi_2(\cdot)$. The prior densities $\pi_j(\cdot)$ are often assumed continuous on the support of $[0, 1]$ for $j = 1, 2$. By Theorem 1 in Diaconis and Ylvisaker (1985), any continuous prior density $\pi(\theta)$ on [0,1] can be uniformly approximated by Bernstein polynomial $\sum_{i=0}^{m} \pi(i/m)\binom{m}{i}\theta^i(1-\theta)^{m-i}$ or equivalently $\sum_{i=0}^{m} w_{im} f_\beta(\theta; i+1, m-i+1)$, where $f_\beta(\cdot; a, b)$ denotes the probability density function of the Beta distribution with parameters a and b and $w_{im} = \pi(\frac{i}{m})/\sum_{i=0}^{m}\pi(\frac{i}{m})$. Therefore, a set of conjugate priors based on mixtures of

Beta densities can be assumed to take the following forms:

$$\begin{aligned}\theta_1 &\sim \sum_{i=0}^{m} w_{1i} f_\beta(\theta_1; i+1, m-i+1) \text{ and} \\ \theta_2 &\sim \sum_{j=0}^{m} w_{2j} f_\beta(\theta_2; j+1, m-j+1),\end{aligned} \tag{1.2}$$

where $w_{lj} \geq 0$ are weights to be determined for $j = 1, 2, ..., m$ subject to the constraint $\sum_{i=0}^{m} w_{li} = 1$ for $l = 1, 2$. Note that the order of Bernstein polynomials can be chosen to different but for simplicity we assume the same order m for both θ_1 and θ_2.

One of the remarkable computational advantages of the mixture of Beta densities is that the prior probability of the null (or alternative) hypothesis can be written as a quadratic form

$$\begin{aligned}& Pr[(\theta_1, \theta_2) \in \Theta_0 | \boldsymbol{w_1}, \boldsymbol{w_2}] \\ = & \int_0^1 \int_0^{g(\theta_1,\rho)} \sum_{j=0}^{m} w_{2j} f_\beta(\theta_2; j+1, m-j+1) d\theta_2 \sum_{i=0}^{m} w_{1i} f_\beta(\theta_1; i+1, m-i+1) d\theta_1 \\ = & \sum_{i=0}^{m} \sum_{j=0}^{m} w_{1i} w_{2j} \left[\int_0^1 \left\{ \int_0^{g(\theta_1,\rho)} f_\beta(\theta_2; j+1, m-j+1) d\theta_2 \right\} f_\beta(\theta_1; i+1, m-i+1) d\theta_1 \right] \\ = & \boldsymbol{w}_1^T A \boldsymbol{w}_2,\end{aligned} \tag{1.3}$$

where $\boldsymbol{w}_1 = (w_{10}, w_{11}, \ldots, w_{1m})^T$, $\boldsymbol{w}_2 = (w_{20}, w_{21}, \ldots, w_{2m})^T$, $A = A(\rho)$ is a $m+1$ by $m+1$ matrix with its element on the pth row and the qth column given by

$$a_{pq}(\rho) = \int_0^1 [F_\beta(g(\theta_1, \rho); q, m-q+2)] f_\beta(\theta_1; p, m-p+2) d\theta_1,$$

and $F_\beta(\cdot; a, b)$ denotes the cumulative distribution function of the $Beta(a, b)$ distribution. Note that here we use a different set of subscripts p and q for the row and column indices because the original subscripts i and j start from 0. The matrix A can be computed using numerical integration (e.g., Gaussian quadratures) for any given integer m. For a given m, the shape parameters in the mixture components of the Beta densities are fixed at the knots $(i-1)/m$

for $i = 1, 2, ..., m, m+1$, so the characteristics of prior density of θ_1 and θ_2 are determined by the weights (w_{li}'s). Next we consider the determination of the prior weights.

Non-informative Prior

When there is no prior information available or the prior information is not reliable, one can choose the weights that balance the prior probabilities of the null and alternative hypotheses. In this sense we call our choice of priors as non-informative (to the selection of hypotheses). Also, for simplicity we let $\boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{w} = (w_0, w_1, \ldots, w_m)^T$ in this case. Consequently, we can choose the weights to minimize the quantity

$$\begin{aligned} \Delta(\boldsymbol{w}, \rho) &= |Pr[(\theta_1, \theta_2) \in \Theta_0 | \boldsymbol{w}] - Pr[(\theta_1, \theta_2) \in \Theta_1 | \boldsymbol{w}]| \\ &= |\boldsymbol{w}^T A \boldsymbol{w} - 0.5| \end{aligned} \tag{1.4}$$

subject to the following restrictions:
i) prior is proper: $w_i \geq 0$ and $\sum_{i=0}^{m} w_i = 1$; and
ii) the weights are symmetric: $w_i = w_{m-i}$ for $i = 0, 1, \ldots, m-1, m$.
Notice that a $\hat{\boldsymbol{w}} = \hat{\boldsymbol{w}}(\rho)$ that satisfies the above requirements for any given matrix $A = A(\rho)$ always exists and can be conveniently expressed as $\hat{\boldsymbol{w}} = \arg\min\{\Delta(\boldsymbol{w}, \rho); \boldsymbol{w}$ satisfies i) and ii)$\}$.

Informative Prior

The control arm in noninferiority trials is the standard treatment that is already available in the market and has established efficacy, therefore we often have some historical data from previous placebo-controlled trial that can be used to construct the prior densities. Suppose we have the estimate θ_1^* (with standard error σ_1^*) for the standard treatment and the estimate θ_0^* (with standard error σ_0^*) for placebo from the historical placebo-controlled trial, now we want to incorporate these into our priors for θ_1 and θ_2 in the noninferiority trial.

First, we determine the weights for the mixture prior density of θ_1 of the form (1.2). Notice that for each component in the mixture of Beta densities with shape-parameters $i+1$ and

$m-i+1$, the mode is at $\frac{i}{m}$. If we measure the distance from the mode of each distribution to the historical estimate θ_1^* by

$$d_i = |\frac{i}{m} - \theta_1^*|, i = 0, 1, 2, \ldots, m-1, m,$$

we can determine the weights $\{w_i\}$ in the mixture prior density according to distances $\{d_i\}$. The general rule of thumb is to give more weights to small d_i's. There could be many ways of calculating the weights to satisfy such a rule. For example, we can let

$$w_{1i} = \frac{1-d_i}{\sum_i (1-d_i)}, i = 0, 1, 2, \ldots, m-1, m.$$

Additionally, the uncertainty of the historical estimate in terms of standard error σ_1^* could be also taken into consideration for prior specification. Generally, if the σ_1^* is small (the historical estimate is relatively accurate), more weights should concentrate near the historical estimate θ_1^*. One easy implementation is to assign positive weights only if the modes of the individual Beta densities in the mixture prior fall into the interval $[\theta_1^* - \sigma_1^*, \theta_1^* + \sigma_1^*]$.

For the prior of θ_2, we may not have historical estimate anymore. We can either use a non-informative flat prior or use a "skeptical" prior based on the historical estimate for placebo θ_0^*, that is,

$$\begin{aligned} \theta_2 &\sim \sum_{j=0}^{m} w_{2j} f_\beta(\theta_2; j+1, m-j+1), \text{ where} \\ w_{2j} &= \frac{1-|\frac{j}{m} - \theta_0^*|}{\sum_j (1-|\frac{j}{m} - \theta_0^*|)}, j = 0, 1, 2, \ldots, m-1, m. \end{aligned}$$

Under this prior, the success rate of the experimental treatment is considered close to that of the placebo before the data is collected. Since the active control has already been shown to be superior to placebo, this prior specification will strongly favor the standard treatment. However, such specification of the prior for θ_2 may be too conservative for testing noninferiority purposes in the situations where the standard treatment is more efficacious than the placebo

by a large margin. Alternatively, once $\boldsymbol{w}_1$ is determined from the historical data, we can also obtain $\boldsymbol{w}_2$ that minimizes $|Pr[(\theta_1, \theta_2) \in \Theta_0|\boldsymbol{w}_2] - Pr[(\theta_1, \theta_2) \in \Theta_1|\boldsymbol{w}_2]|$.

1.2.2 Posterior Inference

Once the priors are determined (e.g., based on the methodologies described in Sections 1.2.1 and 1.2.1), the posterior densities are obtained by using the conjugacy of the mixture of Beta priors:

$$\begin{aligned}\pi(\theta_1|x_1) &= \sum_{i=0}^{m} w_{1i}^* f_\beta(\theta_1; x_1 + i + 1, m + n_1 - x_1 - i + 1) \text{ and} \\ \pi(\theta_2|x_2) &= \sum_{j=0}^{m} w_{2j}^* f_\beta(\theta_2; x_2 + j + 1, m + n_2 - x_2 - j + 1)\end{aligned}$$

with weights

$$w_{1i}^* = \frac{w_{1i}\frac{m+1}{m+n_1+1}\frac{\binom{n_1}{x_1}\binom{m}{i}}{\binom{m+n}{x_1+i}}}{\sum_{i=0}^{m} w_i \frac{m+1}{m+n_1+1}\frac{\binom{n_1}{x_1}\binom{m}{i}}{\binom{m+n_1}{x_1+i}}} \text{ and } w_{2j}^* = \frac{w_{2j}\frac{m+1}{m+n_2+1}\frac{\binom{n_2}{x_2}\binom{m}{j}}{\binom{m+n}{x_2+j}}}{\sum_{j=0}^{m} w_j \frac{m+1}{m+n_2+1}\frac{\binom{n_2}{x_2}\binom{m}{j}}{\binom{m+n_2}{x_2+j}}}.$$

Hence, it follows by a similar argument as in (3), that the posterior probability of the null hypothesis can be written as:

$$Pr[(\theta_1, \theta_2) \in \Theta_0|X_1 = x_1, X_2 = x_2] = \boldsymbol{w}_{\boldsymbol{1}}^{*T} H \boldsymbol{w}_{\boldsymbol{2}}^*, \tag{1.5}$$

where $\boldsymbol{w}_1^* = (w_{10}^*, w_{11}^*, \ldots, w_{1m}^*)^T$, $\boldsymbol{w}_2^* = (w_{20}^*, w_{21}^*, \ldots, w_{2m}^*)^T$, $H = H(x_1, x_2)$ is the $m+1$ by $m+1$ matrix with its element on the pth row and the qth column given by

$$h_{pq}(\rho) = \int_0^1 [F_\beta(g(\theta_1, \rho); x_2 + q, m + n_2 - x_2 - q + 2)] f_\beta(\theta_1; x_1 + p, m + n_1 - x_1 - p + 2) d\theta_1.$$

This is again a quadratic form and only involves one dimensional integration to compute the matrix H, which can be easily accomplished by using any efficient numerical methods (e.g.,

"integrate" function in the software R).

1.2.3 The cut-off value of Bayes factor

For our noninferiority testing problem in particular, the test statistic Bayes factor can be written as

$$\begin{aligned} BF(x_1,x_2) &= \left\{\frac{Pr[(\theta_1,\theta_2)\in\Theta_1|X_1=x_1,X_2=x_2]}{Pr[(\theta_1,\theta_2)\in\Theta_0|X_1=x_1,X_2=x_2]}\right\} / \left\{\frac{Pr[(\theta_1,\theta_2)\in\Theta_1]}{Pr[(\theta_1,\theta_2)\in\Theta_0]}\right\} \\ &= \frac{\boldsymbol{w_1^{*T}} H \boldsymbol{w_2^*}}{1-\boldsymbol{w_1^{*T}} H \boldsymbol{w_2^*}} \cdot \frac{1-\boldsymbol{w_1^T} A \boldsymbol{w_2}}{\boldsymbol{w_1^T} A \boldsymbol{w_2}}, \end{aligned} \tag{1.6}$$

Notice that if we use the non-informative prior the Bayes factor $BF(x_1,x_2) \approx \frac{\boldsymbol{w_1^{*T}} H \boldsymbol{w_2^*}}{1-\boldsymbol{w_1^{*T}} H \boldsymbol{w_2^*}}$.

The Bayes factor is a non-negative function of x_1 and x_2, and its distribution is determined by the marginal distribution of (X_1, X_2), denoted by

$$\begin{aligned} P(x_1,x_2) &= Pr(X_1=x_1,X_2=x_2) \\ &= \sum_{i=0}^{m}\sum_{j=0}^{m} w_i w_j \binom{n_1}{x_1}\binom{n_2}{x_2} * \frac{B(x_1+i+1, n_1-x_1+m-i+1)}{B(i+1,m-i+1)} \\ &\quad * \frac{B(x_2+j+1, n_2-x_2+m-j+1)}{B(j+1,m-j+1)}, \end{aligned} \tag{1.7}$$

with $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, and $\Gamma(\cdot)$ denotes the Gamma function. Thus, it follows that Bayes factor and its marginal distribution can be obtained in closed forms (requiring only one dimensional integration). The one-dimensional integrations required to compute Bayes factor can be accomplished very efficiently using Gaussian quadrature rules.

Since large values of $BF(x_1,x_2)$ are regarded as strong evidence against the null hypothesis, the decision rule can be constructed as rejecting the null hypothesis if $BF(x_1,x_2) > B_0$ for some cut-off value $B_0 > 0$. Suppose $BF(X)$ is the Bayes factor calculated from the observed data, Weiss (1997) suggested one could choose the cut-off value B_0 such that the Bayesian type I error $P[BF(X) > B_0|H_0]$ is less than some pre-specified level α. For the reasons discussed in the previous section, however, we propose determining the cut-off value B_0 that minimizes the total

weighted error criterion that controls both types of errors in a hypothesis testing framework. In this paper we study two versions of such criterion. The first one is based on extensions of usual frequentist type I and II errors as in Weiss (1997). We define the total weighted error as

$$TWE_1(B_0) = (1-\zeta)Pr[BF(X_1, X_2) > B_0|(\theta_1, \theta_2) \in \Theta_0]+\zeta Pr[BF(X_1, X_2) \leq B_0|(\theta_1, \theta_2) \in \Theta_1], \tag{1.8}$$

where $\zeta \in (0, 1)$ is a pre-specified weight that controls relative importance of average type I and type II errors. The Bayesian type I and II error probabilities of the proposed test can be expressed as

$$Pr[BF(X_1, X_2) > B_0|(\theta_1, \theta_2) \in \Theta_0]$$
$$= \frac{\sum_{x_2=0}^{n_2}\sum_{x_1=0}^{n_1} I_{(B_0,\infty)}(BF(x_1,x_2))Pr[(\theta_1,\theta_2)\in\Theta_0|x_1,x_2]P(x_1,x_2)}{Pr[(\theta_1,\theta_2)\in\Theta_0]}$$

and

$$Pr[BF(X_1, X_2) \leq B_0|(\theta_1, \theta_2) \in \Theta_1]$$
$$= \frac{\sum_{x_2=0}^{n_2}\sum_{x_1=0}^{n_1} I_{(0,B_0)}(BF(x_1,x_2))(1-Pr[(\theta_1,\theta_2)\in\Theta_0|x_1,x_2])P(x_1,x_2)}{1-Pr[(\theta_1,\theta_2)\in\Theta_0]},$$

where $I_C(x)$ denotes the indicator function which equals to 1 if $x \in C$ and 0 otherwise.

The extensions of frequestist type I and II errors defined above are conditional on the hypotheses. Lee and Zelen (2000) argued that it is more appropriate to condition on the outcome of the test for the decision errors. These concepts of decision errors have more Bayesian flavor. Following the concepts of errors as suggested by Lee and Zelen (2000), we define the second version of the total weighted error as

$$TWE_2(B_0) = (1-\zeta)Pr[(\theta_1, \theta_2) \in \Theta_0|BF(X_1, X_2) > B_0]+\zeta Pr[(\theta_1, \theta_2) \in \Theta_1|BF(X_1, X_2) \leq B_0], \tag{1.9}$$

where

$$Pr[(\theta_1,\theta_2)\in\Theta_0|BF(X_1,X_2)>B_0]$$
$$=\frac{\sum_{x_2=0}^{n_2}\sum_{x_1=0}^{n_1}I_{(B_0,\infty)}(BF(x_1,x_2))Pr[(\theta_1,\theta_2)\in\Theta_0|x_1,x_2]P(x_1,x_2)}{\sum_{x_2=0}^{n_2}\sum_{x_1=0}^{n_1}I_{(B_0,\infty)}(BF(x_1,x_2))P(x_1,x_2)}$$

and

$$Pr[(\theta_1,\theta_2)\in\Theta_1|BF(X_1,X_2)\leq B_0]$$
$$=\frac{\sum_{x_2=0}^{n_2}\sum_{x_1=0}^{n_1}I_{(0,B_0)}(BF(x_1,x_2))(1-Pr[(\theta_1,\theta_2)\in\Theta_0|x_1,x_2])P(x_1,x_2)}{\sum_{x_2=0}^{n_2}\sum_{x_1=0}^{n_1}I_{(0,B_0)}(BF(x_1,x_2))P(x_1,x_2)}.$$

$BF(X_1,X_2)$ is a function of discrete random variables X_1 and X_2, so it is also a discrete random variable. Therefore, given the sample sizes n_1 and n_2 one can perform an exhaustive search through all possible ordered values of $BF(X_1,X_2)$ to find a cut-off B_0 that minimizes either version of the total weighted error. As a side note, it is more convenient to work on the log scale of the Bayes factor. Since the set $\{(x_1,x_2):BF(x_1,x_2)>B_0\}$ is same as that of $\{(x_1,x_2):\log(BF(x_1,x_2))>L_0\}$ with $L_0=\log(B_0)$, the testing procedure described above remains the same in the logarithmic scale. When using $TWE_1(B_0,\zeta)$ as the criterion it can be shown that $B_0=(1-\zeta)/\zeta$ minimizes TWE_1. Furthermore, Theorem 1 given in the Appendix A indicates that the Bayesian type I error is bounded by the weight ζ if ζ is specified less than some $\zeta_0\in(0,1)$ when $TWE_1(B_0,\zeta)$ is used as the minimizing criterion. For the value of ζ_0, readers can refer to *Remark 1* in the Appendix A. This result provides a useful guidance on the choice of the weight ζ. In particular, if α is the desired level of the test set by a regulatory agency, we can set the weight $\zeta=\alpha$ in TWE_1. Codes written in R are available from the authors upon requests to compute the cut-off values using either criteria.

1.3 Numerical Examples

1.3.1 Simulated Data

In this section, we study some frequentist properties of the proposed Bayesian testing procedure using simulated data. We use odds ratio as the dissimilarity measure. Notice that such a choice of using odds ratio is rather arbitrary and is used here only for illustrative purpose because the proposed testing procedure can be used for hypotheses based on any dissimilarity measure of two binomial parameters. Data were generated under the model $X_1|\theta_1 \sim Bin(n_1, \theta_1)$ and $X_2|\theta_2 \sim Bin(n_2, \theta_2)$. True values of θ_1 were chosen to be $\{0.3, 0.5, 0.8\}$, and the test group proportion $\theta_2 = \eta + g(\rho, \theta_1)$, where $g(\rho, \theta_1) = \frac{\rho\theta_1}{1+\rho\theta_1-\theta_1}$ and η is an incremental factor taking values in the range $[-0.2, 0.2]$ with 0.01 increment. Consequently, the negative values of η favour the null hypothesis, while positive values favour alternative hypothesis.

For the noninferiority "margin" ρ, we followed the guidelines given by Ng (2008) to determine its value. Since there are already extensive literature available that have addressed the problem of choosing a noninferiority margin, we refer the readers for further discussion on the choice of ρ to the recent articles by Hung et al. (2005), Chow and Shao (2006), and Ng (2008) among many others. Under the constancy assumption, Ng (2008) suggested $\rho = (odds(\theta_0)/odds(\theta_1))^{\epsilon}$, where θ_0 is the putative placebo effect, $odds(\theta_0) = \theta_0/(1-\theta_0)$, and $(1-\epsilon)$ is the fraction of the effect size of the active control that one desires to preserve for the experimental treatment. In practice, the quantity $odds(\theta_0)/odds(\theta_1)$ is unknown and hence is replaced by its estimate. However, this difficulty does not exist for our simulation studies. If we set θ_0 to be $\frac{\theta_1}{2}$ and $\epsilon = 0.2$ to preserve at least 80% of the standard treatment effect when claiming noninferiority, then the noninferiority margin $\rho = (odds(\theta_0)/odds(\theta_1))^{\epsilon} = 0.6988$ when $\theta_1 = 0.8$. Also, sample size was included as an experimental factor in our simulation study. For simplicity, sizes of two groups were set equal, i.e., $n = n_1 = n_2 \in \{10, 20, 30, 50\}$. For each combination of experimental factors, $10,000$ replicates of the data were generated and then analyzed by the Bayesian method described in Section 1.2 and the frequentist method using the Blackwelder type test statistic

given by

$$Z_{OR}(x_1, x_2) = \frac{log(\hat{\phi}) - log(\rho)}{\hat{\sigma}_{or}}, \tag{1.10}$$

where

$$\log(\hat{\phi}) = \log \frac{(x_2 + 0.5)(n_1 - x_1 + 0.5)}{(x_1 + 0.5)(n_2 - x_2 + 0.5)} \tag{1.11}$$

and

$$\hat{\sigma}^2_{or} = \frac{1}{x_1 + 0.5} + \frac{1}{n_1 - x_1 + 0.5} + \frac{1}{x_2 + 0.5} + \frac{1}{n_2 - x_2 + 0.5}. \tag{1.12}$$

Note that 0.5 is added to avoid empty cells that make the test statistics not well defined (Tu, 1998). The test rejects the null hypothesis if $Z_{OR}(x_1, x_2) > Z_{1-\alpha}$ for a given $\alpha \in (0, 1)$, where $Z_{1-\alpha}$ denotes the $(1 - \alpha) \times 100\%$ percentile of the standard normal distribution.

For the Bayesian testing procedures, both versions of the total weighted error criterion were investigated. For the reasons given by Theorem 1 in the Appendix A, we let the controlling weight be the same as the size of the frequestist test, namely, $\zeta = \alpha$. As in most medical applications, this weight specification highly emphasizes on controlling type I error. Figures 1.1 and 1.2 show the power curves for $\alpha = 0.05$ and $\alpha = 0.1$ when $\theta_1 = 0.8$, respectively. Notice that $\eta \leq 0$ corresponds to type I error rate while $\eta > 0$ corresponds to power value. As indicated by the fact that the power curves cross the intersection of the vertical and horizontal dotted lines in both Figure 1.1 and Figure 1.2, the proposed Bayesian test with the cut-off minimizing TWE_1 (referred to as BF_H in our figures) has comparable frequentist properties in terms of controlling type I error rate relative to the default frequentist test (given in (1.10)). Meanwhile, the proposed Bayesian method (BF_H) improves the statistical power, especially in small samples. When the sample size gets large, say $n = n_1 = n_2 = 50$ per group, the performance of the Bayesian testing procedure (BF_H) becomes very similar to that of the frequentist method. On the other hand, the Bayesian test with the cut-off value obtained by minimizing TWE_2 (referred to as BF_D) appears to be over-conservative for the weight specification $\zeta = \alpha$. The results corresponding to the true values of $\theta_1 = 0.5$ and $\theta_1 = 0.3$ (not shown) are very similar. All patterns mentioned above are consistent across all θ_1 values that

we have studied.

1.3.2 Real Data Application: Streptococcal Pharyngitis Trial

For further demonstration of the proposed method on real data, we also used the data previously analyzed by Wellek (2003) and Siqueira et al. (2008), which were obtained from a single-center, unblinded, phase IV trial for streptococcal pharyngitis (Scaglione, 1990). In the study, patients with documented group A beta–haemolytic streptococcal pharyngitis were randomized to two treatment groups: 250 mg twice daily clarithromycin and 500 mg twice daily erythromycin. Erythromycin is the standard treatment for the condition of interest while the experimental treatment clarithromycin is considered to have better tolerability. So the question of interest is whether clarithromycin is non-inferior to erythromycin in efficacy. In the subset of the study patients who are 65 or younger, 97 in the erythromycin group of size 107 were observed to have symptoms cured or improved, whereas 98 out of 106 patients in the clarithromycin were successfully treated.

Following Wellek (2003) and Siqueira et al. (2008), we chose odd ratio as the dissimilarity measure for testing the hypotheses $H_0 : \frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)} \leq \rho$ vs. $H_a : \frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)} > \rho$, where θ_1 is the success rate for patients receiving erythromycin and θ_2 denotes the success rate in the clarithromycin group. Again, we set the noninferiority margin $\rho = 0.5$ and the size of test $\alpha = 0.025$ following the specification in Wellek (2003) and Siqueira et al. (2008). Similar to the simulation study, we set the controlling weight $\zeta = \alpha$ and use $m = 20$ as the order of Bernstein polynomial. The observed logarithm of Bayes factor is $\log[BF(X_1 = 97, X_2 = 98)] = 3.218$. For the given data, the critical value for logarithm of the Bayes factor minimizing TWE_1 is $L_{0H} = 3.664$ while the corresponding cut-off minimizing TWE_2 is $L_{0D} = 3.833$. Accordingly, we failed to reject the null hypothesis, hence noninferiority can not be claimed for clarithromycin. These results are consistent with the ones obtained by the frequentist method, in which the p-value based on the Blackwelder-type test is 0.029 (test statistics $Z_{OR} = 1.894$ as defined in (1.10)) and is larger than $\alpha = 0.025$.

In the above example we used $m = 20$ as the order of the Bernstein polynomial. It turns out the test statistic (Bayes factor) and its cut-off values are not very sensitive to the choice of the order of the Bernsterin polynomial. Figure 1.3 shows the corresponding values of the Bayes factor and its cut-offs for this particular data for $m \in \{3, 5, 10, 15, 20\}$. Except the case when $m = 3$, the variation of Bayes factor and its cut-offs seems negligible for various choices of m. Moreover, notice that the decision of not rejecting the null hypothesis remains unchanged for all values of m that we have considered.

1.4 Conclusions and Discussions

We have presented a semiparametric Bayesian approach based on Bernstein polynomials for testing noninferiority in trials with binary outcomes. Results of simulation studies indicate that the proposed Bayesian procedure has comparable frequentist properties of controlling type I error relative to the default frequentist test and meanwhile improves the statistical power, especially in small samples.

Compared to previous Bayesian procedures in the similar context, the proposed approach has serval advantages. First of all, the choice of the prior distribution provides a greater level of flexibility through the specification of weights in the Bernstein polynomial. It also makes less assumption on the form of the prior density function compared to previous Bayesian approaches based on simple parametric conjugate priors. When there is no reliable prior information available, it not only maintains relatively objectivity by balancing the prior probability of two competing hypotheses but also accounts for the design parameter that controls the noninferiority boundary of a particular trial. Secondly, the cut-off value of the test statistic (Bayes factor) is determined by simultaneously controlling both type I and type II errors. Thus, the proposed approach is more suitable in highly regulated settings where controlling type I and type II errors is considered as the most crucial operating characteristic. Lastly, the Bayesian approaches described here are constructed for a general form of the hypotheses for testing noninferiority of two binomial proportion parameters. Although we only demonstrate the testing procedure

based on odds ratio in the numerical examples, other possible dissimilarity measures such as risk difference and relative risk can be easily incorporated into this general framework. Besides mathematical considerations, the choice of dissilimarity measure should be built upon the consensus of medical researchers, statisticians, and most importantly regulatory agencies.

Although not demonstrated by our numerical examples, the weights in the prior distributions can also be specified to accommodate prior information such as historical data and expert opinion. Furthermore, the prior elicitation procedure can be easily generalized to incorporate prior information from multiple sources, such as multiple experts or estimates from multiple placebo-controlled trials. In the real applications, especially in the regulatory setting, how to use informative prior is non-trivial. The specification of the prior distribution should be clinically meaningful for the particular setting of each practical problem. This is why we do not attempt to provide a universally applicable informative prior specification. Noninferiority trials have a unique feature that there is always certain amount prior information available in terms of historical data at least for the active control. How to systematically use historical estimates and their uncertainties becomes important yet challenging issue. Few possible specifications for the informative prior given in this paper try to shed some light towards this direction.

In this paper we only discuss the cases with binary endpoints. However, the concepts of TWE_1 and TWE_2 can be straightforwardly extended to non-binary data. In a general framework where the data (vector) $X \sim f(x|\theta)$ and the parameter (vector) $\theta \sim \pi(\theta)$, where $f(x|\theta)$ can be any sampling density and $\pi(\theta)$ can be any prior density, the two types of total weighted errors can be written as $TWE_1(B_0) = (1-\zeta)\int I[BF(X) > B_0]m_0(X)dX + \zeta\int I[BF(X) \leq B_0]m_1(X)dX$ and $TWE_2(B_0) = (1-\zeta)\frac{\int I[BF(X)>B_0]m_0(X)dX \int_{\Theta_0}\pi(\theta)dX}{\int I[BF(X)>B_0]m_0(X)dX} + \zeta\frac{\int I[BF(X)\leq B_0]m_1(X)dX \int_{\Theta_1}\pi(\theta)dX}{\int I[BF(X)\leq B_0]m(X)dX}$, where $m_0(X) = \int_{\Theta_0} f(X|\theta)\pi_0(\theta)d\theta$, $\pi_0(\theta) = \frac{\pi(\theta)I[\theta\in\Theta_0]}{P[\theta\in\Theta_0]}$, $m_1(X) = \int_{\Theta_1} f(X|\theta)\pi_1(\theta)d\theta$, $\pi_1(\theta) = \frac{\pi(\theta)I[\theta\in\Theta_1]}{P[\theta\in\Theta_1]}$, and $m(X) = \int f(X|\theta)\pi(\theta)d\theta$. Moreover, as pointed out by one of the reviewers, the proposed testing procedure can also apply to the superiority trials by letting $g(\theta_1, \rho) = \theta_1$ in the general form of the hypotheses given in (1.1).

One limitation of the proposed method is that there is some ambiguity in the choice of the

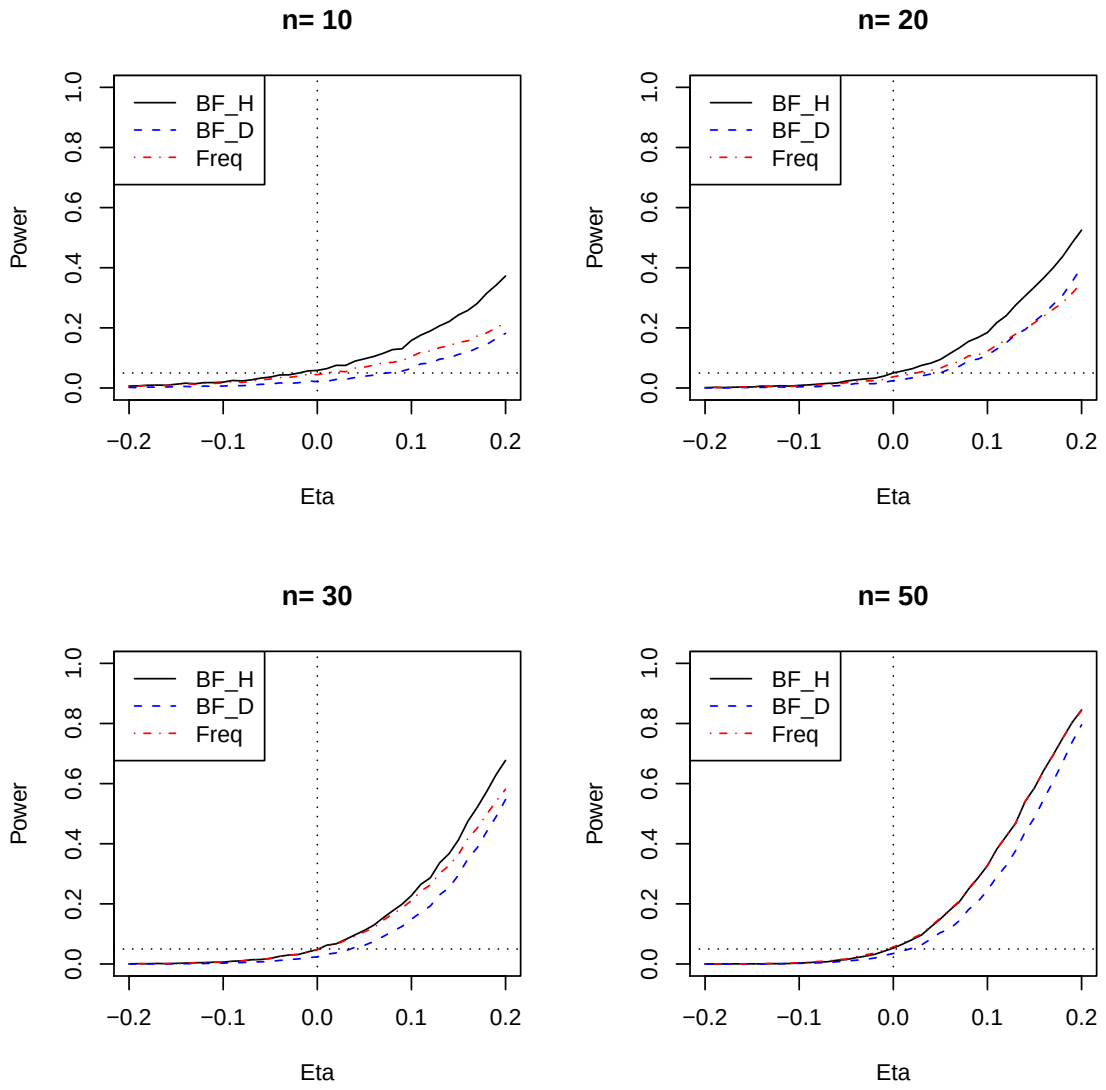
order of Bernstein polynomial. For simplicity, we have used a conservative approach of choosing a large value of m to obtain a reasonably good approximation by using the Weierstrass theorem. More sophisticated approaches such as using data-driven polynomial order can be explored in future studies. Alternatively, some asymptotic results (e.g., $m = o(n^{\frac{2}{5}})$) may also be used in practice (see e.g., Ghosal, 2001; Babu et al., 2002).

Table 1.1: MC Type I error rate and Power of the the proposed Bayesian methods (BF_H and BF_D)and the frequentist approach ($Freq$); $\theta 1 = 0.8, N = 10000, H_0 : odds(\theta_2)/odds(\theta_1) <= \rho, \rho = 0.6988, m = 20$; $\alpha = 0.05$; BF_H (B_0 obtained by minimizing weighted total error contioned on Hypotheses: TWE_1); BF_D (B_0 obtained by minimizing weighted total error contioned on Decisions: TWE_2)

		n (per group)			
η	Method	10	20	30	50
-0.2	BF_H	0.01	0.00	0.00	0.00
	BF_D	0.00	0.00	0.00	0.00
	$Freq$	0.01	0.00	0.00	0.00
-0.15	BF_H	0.01	0.00	0.00	0.00
	BF_D	0.00	0.00	0.00	0.00
	$Freq$	0.01	0.00	0.00	0.00
-0.1	BF_H	0.02	0.01	0.01	0.00
	BF_D	0.01	0.00	0.00	0.00
	$Freq$	0.02	0.01	0.01	0.00
-0.05	BF_H	0.04	0.02	0.02	0.01
	BF_D	0.01	0.01	0.01	0.01
	$Freq$	0.03	0.02	0.02	0.02
0	BF_H	0.06	0.05	0.05	0.05
	BF_D	0.02	0.02	0.02	0.04
	$Freq$	0.04	0.04	0.05	0.06
0.05	BF_H	0.10	0.09	0.11	0.15
	BF_D	0.04	0.05	0.06	0.10
	$Freq$	0.07	0.07	0.11	0.15
0.1	BF_H	0.16	0.18	0.23	0.33
	BF_D	0.07	0.11	0.15	0.25
	$Freq$	0.10	0.12	0.21	0.33
0.15	BF_H	0.24	0.34	0.41	0.59
	BF_D	0.11	0.22	0.30	0.49
	$Freq$	0.15	0.22	0.36	0.59
0.2	BF_H	0.37	0.53	0.68	0.85
	BF_D	0.18	0.40	0.55	0.80
	$Freq$	0.22	0.35	0.58	0.84

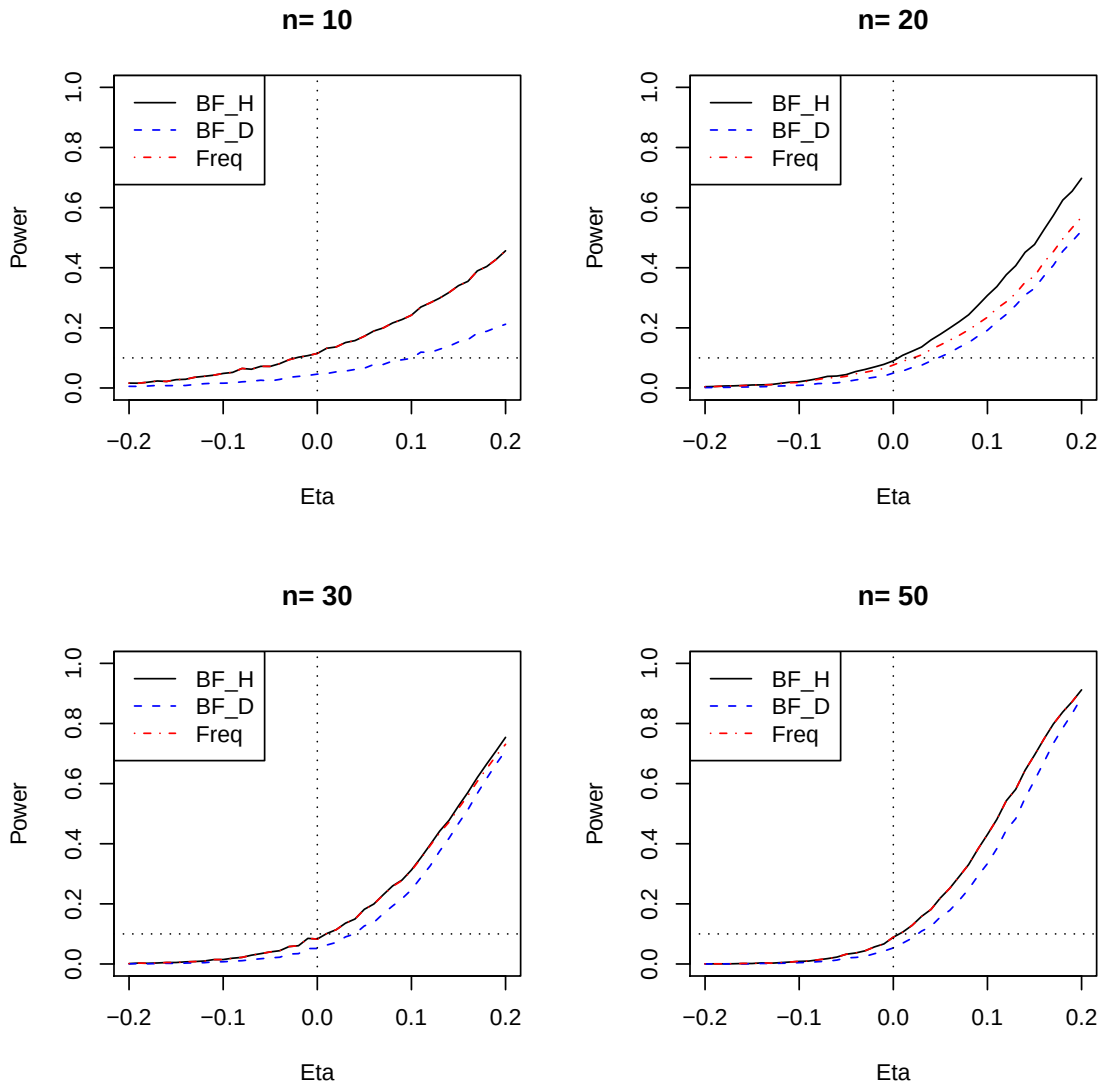
Table 1.2: MC Type I error rate and Power of the the proposed Bayesian methods (BF_H and BF_D)and the frequentist approach ($Freq$); $\theta 1 = 0.8, N = 10000, H_0 : odds(\theta_2)/odds(\theta_1) <= \rho, \rho = 0.6988, m = 20$; $\alpha = 0.10$; BF_H (B_0 obtained by minimizing weighted total error contioned on Hypotheses); BF_D (B_0 obtained by minimizing weighted total error contioned on Decisions)

		n (per group)			
η	Method	10	20	30	50
-0.2	BF_H	0.02	0.00	0.00	0.00
	BF_D	0.01	0.00	0.00	0.00
	$Freq$	0.02	0.00	0.00	0.00
-0.15	BF_H	0.03	0.01	0.01	0.00
	BF_D	0.01	0.00	0.00	0.00
	$Freq$	0.03	0.01	0.01	0.00
-0.1	BF_H	0.05	0.02	0.01	0.01
	BF_D	0.02	0.01	0.01	0.00
	$Freq$	0.05	0.02	0.01	0.01
-0.05	BF_H	0.07	0.04	0.04	0.03
	BF_D	0.02	0.02	0.02	0.02
	$Freq$	0.07	0.04	0.04	0.03
0	BF_H	0.11	0.09	0.08	0.09
	BF_D	0.05	0.05	0.05	0.05
	$Freq$	0.11	0.08	0.08	0.09
0.05	BF_H	0.17	0.18	0.18	0.22
	BF_D	0.07	0.10	0.13	0.16
	$Freq$	0.17	0.14	0.18	0.22
0.1	BF_H	0.24	0.31	0.31	0.43
	BF_D	0.10	0.19	0.25	0.33
	$Freq$	0.24	0.24	0.31	0.43
0.15	BF_H	0.34	0.48	0.53	0.69
	BF_D	0.15	0.33	0.47	0.61
	$Freq$	0.34	0.37	0.52	0.69
0.2	BF_H	0.46	0.70	0.75	0.91
	BF_D	0.21	0.52	0.71	0.89
	$Freq$	0.46	0.57	0.73	0.91



Note: $\theta_1 = 0.8, N = 10000, H_0 : odds(\theta_2)/odds(\theta_1) <= \rho, \rho = 0.6988, m = 20$; $\alpha = 0.05$; BF_H (B_0 obtained by minimizing the weighted total error conditioned on Hypotheses: TWE_1); BF_D (B_0 obtained by minimizing the weighted total error conditioned on Decisions: TWE_2)

Figure 1.1: Monte Carlo Type I error rate and Power of the proposed Bayesian methods (BF_H and BF_D)and the frequentist approach ($Freq$) (the horizontal dotted line denotes the nominal $\alpha = 0.05$ level and the vertical dotted line denotes the boundary of the null and alternative hypotheses)



Note: $\theta_1 = 0.8, N = 10000, H_0 : odds(\theta_2)/odds(\theta_1) <= \rho, \rho = 0.6988, m = 20$; $\alpha = 0.10$; BF_H (B_0 obtained by minimizing the weighted total error conditioned on Hypotheses: TWE_1); BF_D (B_0 obtained by minimizing the weighted total error conditioned on Decisions: TWE_2)

Figure 1.2: Monte Carlo Type I error rate and Power of the proposed Bayesian methods (BF_H and BF_D)and the frequentist approach ($Freq$) (the horizontal dotted line denotes the nominal $\alpha = 0.10$ level and the vertical dotted line denotes the boundary of the null and alternative hypotheses)

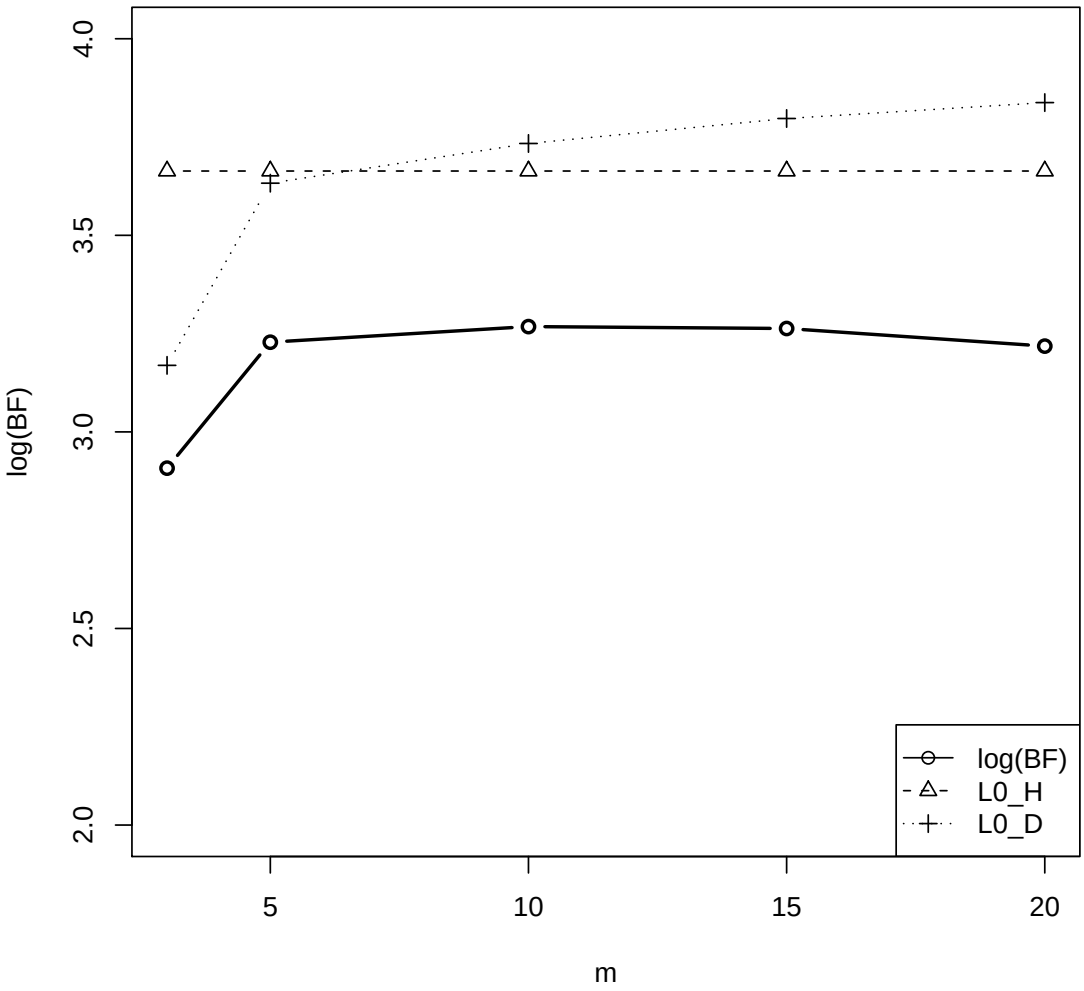


Figure 1.3: Prior sensitivity analysis (for the stretococcal pharyngitis trial) against several values of the order of Bernstein polynomial

Chapter 2

Nonparametric Regression Models for Right-censored Data using Bernstein Polynomials

2.1 Introduction

In regression analysis of survival data, the proportional hazard model (Cox, 1972) has become by far the most widely used method by researchers in many disciplines especially in the field of biostatistics. The most appealing features of the Cox model include the unspecified baseline hazard function and the straightforward interpretation for the effect of categorical covariates such as treatment assignment. In many medical applications, particularly randomized clinical trials, the proportional hazard (PH hereafter) assumption is generally considered reasonable if the study trial does not have a long follow-up time (Perperoglou, Keramopoullos, and van Houwelingen, 2007). In some other situations, however, the validity of this simplification is certainly questionable. It is known that the violation of the PH assumption could lead to erroneous inference in some circumstances (see e.g., Schemper, 1992). In these cases, several alternatives such as proportional odds (PO hereafter) model (Bennett, 1983) and accelerated

failure time (AFT hereafter) model (Kalbfleisch and Prentice, 1980) have been proposed. But these semiparametric models may also turn out to be stringent or even unrealistic in some cases. For instance, consider the case of crossing survival curves, in which there is scientific background to believe that survival curves under different covariate combinations will cross during the study period. For example, in the well known gastric cancer clinical trial (Stablein, Carter and Novak, 1981), patients receiving only chemotherapy may have higher survival rates initially but such rates decay much faster compared to the group of patients receiving chemotherapy and radiotherapy. None of the above models (PH, PO, AFT) can accommodate such a feature of the data.

In the Cox PH model, the conditional hazard function $h(t|Z)$ is modeled as $h(t|Z) = h_0(t)\exp(\beta^T Z)$ where $h_0(\cdot)$ denotes the baseline hazard, Z represents the vector of covariates and the parameter of interest β is constant over time. Several semiparametric extensions of the Cox PH model have been proposed by various authors to relax the proportionality assumption. The most popular approach is to include a time-varying effect $\beta(t)$ by replacing β in the Cox PH model. A challenging step in the time-varying coefficient model is the estimation of the effect function $\beta(\cdot)$. Murphy and Sen (1991) assumed $\beta(\cdot)$ is piecewise constant and proposed a histogram sieve estimator. Zucker and Karr (1990) used smoothing splines based on partial likelihood. The time-varying coefficient model has received extensive attention in literature recently. For further details on more recent approaches on this subject, we refer to the papers by Martinussen and Scheike (2002), Cai and Sun (2003) and Tian, Zucker and Wei (2005). Most of the methodologies involving the time-varying effect may turn out to be computationally intensive because the form of likelihood function is usually very complicated. Besides, Perperoglou et al. (2007) pointed out that when survival curves cross, over-emphasizing on the regression coefficient might not be appropriate. The reason is that the appealing feature of easy interpretation and estimation of the Cox PH model, which comes from the separation of time and covariate effects, will be lost under nonproportional hazards. Instead, the entire conditional hazard or survival curve will be more informative to medical researchers. Therefore,

in this paper we take a different direction by directly modeling the conditional hazard function using Bernstein polynomials.

In contrast to various extensions of the Cox PH model to accommodate nonproportionality, our method is completely nonparametric and computationally much simpler to implement. Fully nonparametric hazard regression models have been studied by many authors, most of which focused on the kernel method and smoothing splines (see e.g., Li and Doss, 1995; Gray, 1996; Spierdijk, 2008, among others). One important methodology in this line of research is referred to as HARE (Hazard Regression) in Kooperberg, Stone, and Truong (1995). HARE is a regression model based on linear splines and their tensor products for the conditional log-hazard function. In the HARE model, linear splines are used rather than quadratic or cubic splines in order to avoid numerical integration in the log-likelihood (and in its gradient and Hessian matrix). This simplification is essential due to the model selection step involving stepwise addition and stepwise deletion incorporated in HARE. However, in some situations where the conditional log-hazard function takes complex form the linear splines and their tensor products may not capture the overall dependence of the event time on other covariates.

Bernstein polynomials have been considered in a wide range of statistical problems based on completely observed data. The most common application is density estimation, which dates back to the work of Vitale (1975). Some of the most recent work on this topic includes Petrone (1999), Babu, Canty, and Chaubey (2002), Choudhuri, Ghosal, and Roy (2004) among many others. Bernstein polynomials have also been applied in the regression setting by Tenbusch (1994) and Chang et al. (2007). In the context of survival analysis with censored data, Chang et al. (2005) proposed using Bernstein polynomials for hazard rate estimation in a Bayesian framework for a homogeneous population, i.e., without any covariates.

In this paper, we consider nonparametric hazard regression based on Bernstein polynomials for right-censored data. As we will demonstrate later, Bernstein polynomials have several advantages in this particular setting. Monotonicity of the cumulative hazard function can be modeled naturally via Bernstein polynomials. In addition, Bernstein polynomials have nice

differentiability properties such that the log-likelihood, its gradient, and Hessian matrix all take relatively easy forms, making our method very easy to implement as compared to other computationally intensive methods such as those based on the time-varying coefficient models. To obtain a smooth estimator for the conditional hazard function in a full nonparametric setting, we use a sieve maximum likelihood estimator (Grenander, 1981; Geman and Hwang, 1982). The proposed nonparametric regression model in this paper is shown to nest the Cox PH model as a special case. The rest of the paper proceeds as follows. In section 3.2, we describe the model for categorical and continuous covariates. We show that the sieve maximum likelihood estimate is consistent and the corresponding rate of convergence is derived in section 2.3. In section 2.4, the proposed method is demonstrated through simulated data as well as real data from two well known cancer studies. Finally, we conclude with some discussions in section 2.5.

2.2 Conditional Hazard Model using Bernstein Polynomials

First, we consider the one-sample right-censored data with no covariates. Suppose an experiment or a clinical trial consists of n subjects, T_i denotes the time to certain event of interest for the subject i and assume that T_i's are independent and identically distributed (iid.) for $i = 1, 2, ..., n$. The event time T_i is subject to random right-censoring C_i and hence for each subject we observe (X_i, Δ_i), where $X_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, and $I(A)$ denotes the indicator function that takes the value 1 when the event A is true, otherwise $I(A) = 0$. We also assume that T_i is statistically independent of C_i for each $i = 1, 2, ..., n$. For any $t \geq 0$, the cumulative hazard function is given by $H(t) = -\log S(t)$ and the hazard function $h(t) = \dot{H}(t)$, where $S(t) = Pr(T_i > t)$ is the survival function and $\dot{H}(t)$ denotes the derivative of $H(t)$. Further, following standard practice (e.g. Tian et al., 2005) we assume that there exists a $\tau < \infty$ such that $\tau = \inf\{t : S(t) = 0\}$.

Now we define the model for the hazard function as

$$h_m(t,\boldsymbol{\gamma}) = \sum_{k=1}^{m} \gamma_k g_{m,k}(t) = \boldsymbol{\gamma}^T \boldsymbol{g}_m(t),\ 0 \le t < \infty, \tag{2.1}$$

where the coefficients $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_m)^T$ satisfy $\gamma_k \ge 0$ for all k's and $\boldsymbol{g}_m(t) = (g_{m,0}(t), g_{m,1}(t), ..., g_{m,m}(t))^T$ with the base functions satisfying $g_{m,k}(\cdot) \ge 0$ and $\int_0^\infty g_{m,k}(u)du < \infty$ for all $k \le m$. Therefore, we have the corresponding model for the cumulative hazard function

$$H_m(t,\boldsymbol{\gamma}) = \sum_{k=1}^{m} \gamma_k G_{m,k}(t) = \boldsymbol{\gamma}^T \boldsymbol{G}_m(t),\ 0 \le t < \infty, \tag{2.2}$$

where $\boldsymbol{G}_m(t) = (G_{m,0}(t), G_{m,1}(t), ..., G_{m,m}(t))^T$ with $G_{m,k}(t) = \int_0^t g_{m,k}(u)du$. Clearly, the monotonicity of $H_m(\cdot)$ is enforced by the restriction that $\gamma_k \ge 0$ and $g_{m,k}(\cdot) \ge 0$ for $k = 1, 2, ..., m$. Although the above model for $H_m(\cdot)$ does not satisfy the requirement that $H_m(\tau) = \infty$, a simple tail adjustment can be made to satisfy the requirement (see section 2.2.4 for further details).

In this paper, we construct the base functions $g_{m,k}(\cdot)$ and $G_{m,k}(\cdot)$ using Bernstein polynomials, which according to a result by Carciner and Peña (1993) has the best shape-preserving property among all approximation polynomials. For a continuous function such as a cumulative hazard $H(\cdot)$ on $(0, \tau]$, the approximating Bernstein polynomial of order m is given by

$$B(t; m, H) = \sum_{k=0}^{m} H\left(\frac{k}{m}\tau\right) \binom{m}{k} (t/\tau)^k (1 - t/\tau)^{m-k}.$$

By the Weierstrass theorem, $B(\cdot; m, H) \to H(\cdot)$ uniformly over $(0, \tau]$ as $m \to \infty$ (Lorentz,

1956). Also, the derivative of the Bernstein polynomial for H can be written as

$$\begin{aligned}\dot{B}(t;m,H) &= \frac{\partial B(t;m,H)}{\partial t} \\ &= \sum_{k=1}^{m}\left\{H\left(\frac{k}{m}\tau\right) - H\left(\frac{k-1}{m}\tau\right)\right\}\frac{f_\beta(t/\tau;k,m-k+1)}{\tau},\end{aligned}$$

where $f_\beta(\cdot;k,m-k+1)$ is the probability density function of the Beta distribution with shape parameters k and $m-k+1$. It is also well known that $\dot{B}(\cdot;m,H) \to h(\cdot)$ uniformly on $(0,\tau]$ as $m \to \infty$ (Lorentz, 1956).

Now we let the base function $g_{m,k}(t) = f_\beta(t/\tau;k,m-k+1)/\tau$ in (2.1), so

$$G_{m,k}(t) = \int_0^t g_{m,k}(u)du = \int_0^t f_\beta(u/\tau;k,m-k+1)d(u/\tau),$$

i.e., $G_{m,k}(t)$ becomes the cumulative distribution function of the Beta distribution with shape parameters k and $m-k+1$ evaluated at t/τ. Consequently, the log-likelihood function of γ corresponding to the models (2.1) and (2.2) can be written as

$$\begin{aligned}l(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\{\Delta_i \log(h_m(X_i,\boldsymbol{\gamma})) - H_m(X_i,\boldsymbol{\gamma})\} \\ &= \sum_{i=1}^{n}\{\Delta_i \log(U_i^T\boldsymbol{\gamma}) - V_i^T\boldsymbol{\gamma}\}, \qquad (2.3)\end{aligned}$$

where $\boldsymbol{\gamma} \in \mathcal{C}_m = [0,\infty)^m$, $U_i = \boldsymbol{g}_m(X_i)$, and $V_i = \boldsymbol{G}_m(X_i)$.

Remark 1: Since $l(\boldsymbol{\gamma})$ as defined in (2.3) is a strictly concave function and $l(\boldsymbol{\gamma}) \to -\infty$ as $\boldsymbol{\gamma} \to \partial\mathcal{C}_m$, where $\partial\mathcal{C}_m$ denotes the boundary of $\mathcal{C}_m$, the existence and uniqueness of the maximum likelihood estimator $\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}\in\mathcal{C}_m} l(\boldsymbol{\gamma})$ follow immediately.

2.2.1 Categorical Covariate

For simplicity, first we consider a dichotomous covariate, say $Z \in \{0,1\}$. For example, in most medical applications Z denotes the assignment of the treatment group: 0 for the placebo and

1 for the active treatment. Essentially, the model for the hazard function and the cumulative hazard function can be easily extended by using different sets of parameter vectors for different groups as it is in (2.1) and (2.2), namely,

$$\begin{aligned} h_m(t, \boldsymbol{\gamma}|Z) &= \{(1-Z)\boldsymbol{\gamma}_{\mathbf{0}}^T + Z\boldsymbol{\gamma}_{\mathbf{1}}^T\}\boldsymbol{g}_m(t) \text{ and} \\ H_m(t, \boldsymbol{\gamma}|Z) &= \{(1-Z)\boldsymbol{\gamma}_{\mathbf{0}}^T + Z\boldsymbol{\gamma}_{\mathbf{1}}^T\}\boldsymbol{G}_m(t), \end{aligned} \tag{2.4}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\mathbf{0}}^T, \boldsymbol{\gamma}_{\mathbf{1}}^T)^T = (\gamma_{01}, \gamma_{02}, ..., \gamma_{0m}, \gamma_{11}, \gamma_{12}, ..., \gamma_{1m})^T$, $g_{m,k}(\cdot)$ and $G_{m,k}(\cdot)$ are the same as they were defined in the previous section. Accordingly, the log-likelihood function becomes

$$\begin{aligned} l(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\{\Delta_i \log(h_m(X_i, \boldsymbol{\gamma}|Z_i)) - H_m(X_i, \boldsymbol{\gamma}|Z_i)\} \\ &= \sum_{i=1}^{n}\{\Delta_i \log(U_i^T\boldsymbol{\gamma}) - V_i^T\boldsymbol{\gamma}\}, \end{aligned} \tag{2.5}$$

where

$$U_i = \begin{bmatrix} (1-Z_i)\boldsymbol{g}_m(X_i) \\ Z_i\boldsymbol{g}_m(X_i) \end{bmatrix} \text{ and } V_i = \begin{bmatrix} (1-Z_i)\boldsymbol{G}_m(X_i) \\ Z_i\boldsymbol{G}_m(X_i) \end{bmatrix}.$$

The log-likelihood function in (2.5) is of the same form as in the case of one-sample data with no covariates (see eq. (2.3)), so the existence and uniqueness of the maximum likelihood estimator still hold in the presence of a binary covariate.

The model described in (2.4) can be regarded as modeling the discretized hazard function using 1-way ANOVA. As a result, it can be further extended to the cases when there are multiple categorical covariates and each has more than 2 levels. Generally, multiple categorical covarites can be summarized in terms of one categorical covarite taking multiple levels that index all the possible combinations of different levels in multiple categorical covariates. Suppose $Z \in \{1, 2, ..., J\}$ summarizes all categorical covariates, then the model in (2.4) can be expressed

as

$$\begin{aligned} h_m(t,\boldsymbol{\gamma}|Z) &= \left\{\sum_{j=1}^{J} I(Z=j)\boldsymbol{\gamma_j}^T\right\}\boldsymbol{g}_m(t) \text{ and} \\ H_m(t,\boldsymbol{\gamma}|Z) &= \left\{\sum_{j=1}^{J} I(Z=j)\boldsymbol{\gamma_j}^T\right\}\boldsymbol{G}_m(t), \end{aligned} \tag{2.6}$$

where $\boldsymbol{\gamma_j} = (\gamma_{j1}, \gamma_{j2}, ..., \gamma_{jm})^T$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma_1}^T, \boldsymbol{\gamma_2}^T, ..., \boldsymbol{\gamma_J}^T)^T$. Again, the corresponding log-likelihood function can be written in the form as in (2.5) and hence the existence and uniqueness of maximum likelihood estimate $\hat{\boldsymbol{\gamma}}$ follow by the *Remark 1* stated earlier.

2.2.2 Continuous Covariate

Next we consider the case when Z is a continuous convariate. As natural extensions of (2.1) and (2.2), the models for the hazard function and the cumulative hazard function can be written as

$$\begin{aligned} h_m(t,\boldsymbol{\gamma}|Z) &= \sum_{k=1}^{m} \gamma_k(Z) g_{m,k}(t) = \boldsymbol{\gamma}(Z)^T \boldsymbol{g}_m(t) \text{ and} \\ H_m(t,\boldsymbol{\gamma}|Z) &= \sum_{k=1}^{m} \gamma_k(Z) G_{m,k}(t) = \boldsymbol{\gamma}(Z)^T \boldsymbol{G}_m(t). \end{aligned} \tag{2.7}$$

One possible approach of conditioning on a continuous covariate is to discretize it into several categories, then it could be fitted into the framework described in the previous section. In order to capture the overall dependence between Z and the response (via the conditional hazard or cumulative hazard functions), it may require a relatively large number of discretized categories. As a consequence, it may lead to small number of samples within each category. Instead of modeling $\gamma_k(Z)$ as a linear function of Z, we propose approximating the function $\gamma_k(\cdot)$ with another $\tilde{m}$-th order Bernstein polynomial. To implement this, a transformation $\tilde{Z} = a(Z)$ is required to map the continuous covariate to the unit interval $(0, 1)$. Then the second fold of Bernstein polynomial can be written as $\gamma_k(Z) = \sum_{j=1}^{\tilde{m}} \gamma_{kj} \binom{\tilde{m}}{j} \tilde{Z}^j (1-\tilde{Z})^{\tilde{m}-j}$. If

we let $\tilde{m} = m - 1$, the expression can be rewritten as

$$\gamma_k(Z) = \sum_{j=1}^{m} \gamma_{kj} \tilde{g}_{m,j}(\tilde{Z}) = \boldsymbol{\gamma}_k^T \tilde{\boldsymbol{g}}_m(\tilde{Z}),$$

where $\tilde{g}_{m,j}(\tilde{Z}) = f_\beta(\tilde{Z}, j, m - j + 1)/m$. Note that even though $\tilde{g}_{m,j}(\cdot)$ and $g_{m,j}(\cdot)$ take very similar forms, they are not exactly identical because the function $g_{m,j}(\cdot)$ involves the cut-off time point τ. We can now express the models as

$$\begin{aligned} h_m(t, \boldsymbol{\gamma}|Z) &= \sum_{k=1}^{m}\sum_{j=1}^{m} \gamma_{kj} \tilde{g}_{m,j}(\tilde{Z}) g_{m,k}(t) \text{ and} \\ H_m(t, \boldsymbol{\gamma}|Z) &= \sum_{k=1}^{m}\sum_{j=1}^{m} \gamma_{kj} \tilde{g}_{m,j}(\tilde{Z}) G_{m,k}(t), \end{aligned} \tag{2.8}$$

where the unknown parameter $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T, ..., \boldsymbol{\gamma}_m^T)^T$ with $\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2}, ..., \gamma_{km})^T$ for $k = 1, 2, ..., m$. Notice that it is not necessary to use $\tilde{m} = m - 1$, but we find it convenient to implement our method with one tuning parameter m.

Then the log-likelihood function becomes

$$\begin{aligned} l(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} \{\Delta_i \log(h_m(X_i, \boldsymbol{\gamma}|Z_i)) - H_m(X_i, \boldsymbol{\gamma}|Z_i)\} \\ &= \sum_{i=1}^{n} \{\Delta_i \log(U_i^T \boldsymbol{\gamma}) - V_i^T \boldsymbol{\gamma}\}, \end{aligned} \tag{2.9}$$

where $U_i = vec[\tilde{\boldsymbol{g}}_m(\tilde{Z}_i)\boldsymbol{g}_m^T(X_i)]$, $V_i = vec[\tilde{\boldsymbol{g}}_m(\tilde{Z}_i)\boldsymbol{G}_m^T(X_i)]$, and $vec[\cdot]$ denotes the vectorization by column operator applied to a matrix. Similarly, we impose the restriction $\gamma_{kj} > 0$ for $1 \le k, j \le m$ for $m = 1, 2, ...$ Because the above form is the same as in (2.3), it immediately guarantees the existence and uniqueness of maximum likelihood estimate for a given m.

As a special case of our model in (2.8), when the time effect and the covariate effect can be

factored as $\gamma_{kj} = \gamma_k \tilde{\gamma}_j$ in (2.8), the model simplifies to

$$\begin{aligned} h_m(t, \boldsymbol{\gamma}|Z) &= \left\{ \sum_{k=1}^{m} \gamma_k g_{m,k}(t) \right\} \left\{ \sum_{j=1}^{m} \tilde{\gamma}_j \tilde{g}_{m,j}(\tilde{Z}) \right\} = h_{m0}(t) \exp[\mu(Z)] \text{ and} \\ H_m(t, \boldsymbol{\gamma}|Z) &= \left\{ \sum_{k=1}^{m} \gamma_k G_{m,k}(t) \right\} \left\{ \sum_{j=1}^{m} \tilde{\gamma}_j \tilde{g}_{m,j}(\tilde{Z}) \right\} = H_{m0}(t) \exp[\mu(Z)], \end{aligned}$$

which is a proportional hazard model with the baseline hazard approximated by Bernstein polynomials and the covariate effect $\mu(Z) = \log[\sum_{j=1}^{m} \tilde{\gamma}_j \tilde{g}_{m,j}(\tilde{Z})]$. Notice that the Cox proportional hazard model is a special case of this model when $\mu(\cdot)$ is assumed linear in Z.

2.2.3 Categorical and Continuous Covariates

When both a categorical covariate Z_1 and a continuous covariate Z_2 are present, generally we can stratify the data based on the levels of the categorical covariate Z_1 and fit a separate model with the continuous covariate Z_2 at each stratum. This stratified approach is suitable for large sample sizes where there are enough data at each stratum. But for moderate or small datasets, some subgroup may result in only a few samples so that corresponding Z_2's (or transformed version $\tilde{Z}_2$) may not adequately cover the entire unit interval $(0, 1)$. Alternatively, if we are willing to accept the assumption that there is no interaction between the categorical covariate Z_1 and the continuous covariate Z_2, we can write a partially linear coefficient model as

$$\begin{aligned} h_m(t, \gamma|Z_1, Z_2) &= \boldsymbol{\gamma}(Z_1, Z_2)^T \boldsymbol{g}_m(t) = \{\boldsymbol{\phi}(Z_1) + \boldsymbol{\psi}(Z_2)\}^T \boldsymbol{g}_m(t) \text{ and} \\ H_m(t, \gamma|Z_1, Z_2) &= \boldsymbol{\gamma}(Z_1, Z_2)^T \boldsymbol{G}_m(t) = \{\boldsymbol{\phi}(Z_1) + \boldsymbol{\psi}(Z_2)\}^T \boldsymbol{G}_m(t), \end{aligned} \tag{2.10}$$

where $\boldsymbol{\gamma}(Z_1, Z_2) = (\gamma_1(Z_1, Z_2), \gamma_2(Z_1, Z_2), ..., \gamma_m(Z_1, Z_2))^T$ with $\gamma_k(Z_1, Z_2) = \phi_k(Z_1) + \psi_k(Z_2)$, $\phi_k(Z_1)$ is just an ANOVA model with mean effects at the levels of Z_1, and $\psi_k(Z_2)$ is a smooth function. Suppose Z_1 has J levels, then we can write $\boldsymbol{\phi}(Z_1)^T \boldsymbol{g}_m(t) = U_1^T \boldsymbol{\mu}$ and $\boldsymbol{\phi}(Z_1)^T \boldsymbol{G}_m(t) = V_1^T \boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, ..., \boldsymbol{\mu}_J^T)^T$ with $\boldsymbol{\mu}_j = (\mu_{j1}, \mu_{j2}, ..., \mu_{jm})^T$,

$$U_1 = \begin{bmatrix} I(Z_1 = 1)\boldsymbol{g}_m(t) \\ I(Z_1 = 2)\boldsymbol{g}_m(t) \\ \vdots \\ I(Z_1 = J)\boldsymbol{g}_m(t) \end{bmatrix} \text{ and } V_1 = \begin{bmatrix} I(Z_1 = 1)\boldsymbol{G}_m(t) \\ I(Z_1 = 2)\boldsymbol{G}_m(t) \\ \vdots \\ I(Z_1 = J)\boldsymbol{G}_m(t) \end{bmatrix}.$$

Similar to the single continuous covariate case, we can write $\boldsymbol{\psi}(Z_2)^T\boldsymbol{g}_m(t) = U_2^T\boldsymbol{w}$ and $\boldsymbol{\psi}(Z_2)^T\boldsymbol{G}_m(t) = V_2^T\boldsymbol{w}$, where $\boldsymbol{w} = (\boldsymbol{w}_1^T, \boldsymbol{w}_2^T, ..., \boldsymbol{w}_m^T)^T$ with $\boldsymbol{w}_k = (w_{k1}, w_{k2}, ..., w_{km})^T$, $U_2 = vec[\tilde{\boldsymbol{g}}_m(Z_2)\boldsymbol{g}_m^T(t)]$ and $V_2 = vec[\tilde{\boldsymbol{g}}_m(Z_2)\boldsymbol{G}_m^T(t)]$. Therefore, the partially linear coefficient model can be written in a unified form as

$$h_m(t, \boldsymbol{\gamma}|Z_1, Z_2) = U^T\boldsymbol{\gamma} \text{ and } H_m(t, \boldsymbol{\gamma}|Z_1, Z_2) = V^T\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma} = (\boldsymbol{\mu}^T, \boldsymbol{w}^T)^T$, $U = [U_1^T, U_2^T]^T$, and $V = [V_1^T, V_2^T]^T$. The log-likelihood function again takes the form as in (2.3) and hence the *Remark 1* applies to this case as well.

Thus, we have shown that for a broad range of applications including categorical and continuous predictors, the resulting log-likelihood function can be written in the simple form (2.3) as in the case of a one-sample problem. This attractive feature of our method makes our non-parametric approach very easy to implement requiring a simple numerical iterative method to estimate the conditional hazard function without making any stringent restrictive assumptions.

2.2.4 Computational Details

As discussed in the previous sections, the log-likelihood function in all cases can be written in a general form (2.3) as $l(\boldsymbol{\gamma}) = \sum_{i=1}^n \{\Delta_i \log(U_i^T\boldsymbol{\gamma}) - V_i^T\boldsymbol{\gamma}\}$, where U_i and V_i are some vectors depending only on data. Consequently, the gradient and the Hessian matrix of the log-likelihood function can be conveniently expressed as

$$\frac{\partial l}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^n \left(\frac{\Delta_i U_i}{U_i^T\boldsymbol{\gamma}} - V_i\right) \text{ and } \frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T} = -\sum_{i=1}^n \frac{\Delta_i U_i U_i^T}{(U_i^T\boldsymbol{\gamma})^2}. \tag{2.11}$$

Although the gradient of the log-likelihood function has a trackable form, it still seems difficult if not impossible to solve the likelihood equation analytically. However, since the gradient vector and the Hessian matrix are in simple closed forms, one can easily implement Newton-Raphson method or its various extensions to compute the maximum likelihood estimate $\hat{\boldsymbol{\gamma}}$. For example, the maximum can be obtained iteratively through $\hat{\boldsymbol{\gamma}}^{(r+1)} = \hat{\boldsymbol{\gamma}}^{(r)} - [\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\left\{\hat{\boldsymbol{\gamma}}^{(r)}\right\}]^{-1} \frac{\partial l}{\partial \boldsymbol{\gamma}}\left\{\hat{\boldsymbol{\gamma}}^{(r)}\right\}$, where superscript r denotes the value at the r-th iteration starting with say $\hat{\boldsymbol{\gamma}}^{(0)} = \mathbf{1}$. However, for the implementation in this paper we use a modification of the quasi-Newton method given by Byrd et al. (2002), which is available as an option in the R function "optim". Finally, the variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$ can be approximated by inverting the observed information matrix: $\hat{V}(\hat{\boldsymbol{\gamma}}) = [\sum_{i=1}^{n} \frac{\Delta_i U_i U_i^T}{(U_i^T \hat{\boldsymbol{\gamma}})^2}]^{-1}$.

In the previous section we mentioned that a suitable transformation $a(\cdot)$ is needed to map the continuous covariate into the unit interval $(0, 1)$. Instead of a non-linear transformation we choose $a(\cdot)$ such that $\tilde{Z} = a(Z)$ is approximately linear in Z on the observed range of Z's. Although many other transformations are possible, throughout this paper we use

$$\tilde{Z} = a(Z) = \frac{Z - Z_{(1)} + \epsilon}{Z_{(n)} - Z_{(1)} + 2\epsilon},$$

where we let $\epsilon = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2}$, $Z_{(1)} = \min\{Z_1, Z_2, ..., Z_n\}$, and $Z_{(n)} = \max\{Z_1, Z_2, ..., Z_n\}$.

For the value of the cut-off time point, we let $\hat{\tau} = X_{(n)} = \max\{X_1, X_2, ..., X_n\}$ which denotes the largest observed value. As described in the previous sections, we model the hazard function and the cumulative hazard function over $(0, \hat{\tau}]$ using Bernstein polynomials. When $t > \hat{\tau}$, since we do not have any information from data about the distribution of the event time, we simply let the hazard rate stay constant after the cut-off time point $t = \hat{\tau}$. The overall hazard function

over $t \in [0, \infty)$ can be written as

$$h_m^*(t, \boldsymbol{\gamma}|Z) = \begin{cases} h_m(t, \boldsymbol{\gamma}|Z) & \text{, if } 0 \le t < \hat{\tau} \\ \frac{m\gamma_m(Z)}{\tau} & \text{, if } t \ge \hat{\tau}. \end{cases}$$

Correspondingly, the cumulative hazard function over $t \in [0, \infty)$ becomes

$$H_m^*(t, \boldsymbol{\gamma}|Z) = \begin{cases} H_m(t, \boldsymbol{\gamma}|Z) & \text{, if } 0 \le t < \hat{\tau} \\ \sum_{k=1}^m \gamma_k(Z) + \frac{m\gamma_m(Z)}{\tau}(t - \tau) & \text{, if } t \ge \hat{\tau}. \end{cases}$$

Now we have $\int_0^\infty h_m^*(t, \boldsymbol{\gamma}|Z)dt = \infty$ for any $\boldsymbol{\gamma}$ and Z, so the tail adjustment suffices that the overall hazard $h_m^*(\cdot)$ is a legitimate hazard function. On the other hand, since we let $\hat{\tau} = X_{(n)}$, all the data entering into the likelihood are modeled by Bernstein polynomials, hence the properties of the model in the previous and the following sections remain unaffected. As a side note, it is easy to verify that $\hat{\tau}_n \to_p \tau$ as $n \to \infty$ where $\tau = \inf\{t > 0 : S(t) = 0\}$.

2.3 Asymptotic Properties

In this section we present the asymptotic properties of the sieve maximum likelihood estimator for the hazard rate of the event time: $\hat{h}_m(t) = h_m(t, \boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}_{mle})$. For brevity, we only show the consistency and the rate of convergence for the one-sample case. Since the log-likelihood function takes similar forms as in (2.3) even in presence of categorical or continuous covariates, the asymptotic properties verified in the section can be readily extended to the case of regression model if we assume that the observations (X_i, Δ_i, Z_i) are iid and T_i is independent of C_i given Z_i for each $i = 1, 2, ..., n$.

In the context of right-censored data with no covariate, for each subject we observe iid copies of $W = (X, \Delta)$ where $X = \min(T, C)$ for the event (or failure) time T and the random censoring time C and $\Delta = I(T \le C)$. Throughout this article we assume that the distributions

of T and C are dominated by Lebesgue measure. We further assume that the censoring is noninformative, that is, T and C are independent. In the presence of covariate Z we assume that T and C are conditionally independent given Z. We let $h_0(\cdot)$ denote the true hazard function of T and $h_c(\cdot)$ denote the hazard function of C.

Now suppose the data $W_i = (X_i, \Delta_i) \overset{iid}{\sim} P_\theta$ and if we let μ be the product measure of Lebesgue measure and counting measure on $\{0, 1\}$, we have the density of W_i:

$$\begin{aligned} p_\theta(w = (x, \delta)) &= \frac{dP_\theta}{d\mu}(x, \delta) \\ &= \left\{\theta(x)^\delta \exp\left(-\int_0^x \theta(u)du\right)\right\} * \left\{h_c(x)^{1-\delta} \exp\left(-\int_0^x h_c(u)du\right)\right\}, \end{aligned}$$

where the hazard rate of the censoring time $h_c(\cdot)$ can be regarded as a nuisance parameter and as it does not enter into the calculation of the maximum likelihood estimator we can ignore estimating this function. For brevity, we let $f(x, \delta, \theta) = \theta(x)^\delta \exp\{-\int_0^x \theta(u)du\}$ and $g(x, \delta, h_c) = h_c(x)^{1-\delta} \exp\{-\int_0^x h_c(u)du\}$. Finally, $E_0[\cdot]$ and $VAR_0[\cdot]$ denote the expectation and the variance operator with respect to the true density of W: $p_0(w) = f(h_0, x, \delta) * g(h_c, x, \delta)$. We also assume that

(I) $\tau = \inf\{t > 0 : \int_0^t h_0(u)du = \infty\} < \infty$.

We study the consistency and the rate of the convergence using the Hellinger distance as the metric of choice

$$d(\theta_1, \theta_2) = \left\{\int (p_{\theta_1}^{1/2} - p_{\theta_2}^{1/2})^2 d\mu\right\}^{1/2}, \tag{2.12}$$

where $\theta_1 = \theta_1(\cdot)$ and $\theta_2 = \theta_2(\cdot)$ denote two hazard functions in the parameter space Θ to be defined shortly. There are several reasons for using this metric. Although we are estimating the hazard function, a more natural and practical entity of interest to the medical researchers is the comparison of the survival curves, or ultimately the global behavior of the corresponding densities. As pointed out by van de Geer (2000), the Hellinger distance is most convenient in this situation. Additionally, the Hellinger distance has the advantage that it is always well defined for arbitrary continuous density functions.

To study the asymptotic properties of $\hat{h}_m(\cdot)$, the following conditions on the boundness and smoothness of the true hazard function h_0 are assumed.

(II) $h_0(\cdot)$ *is continuous on* $[0,\tau]$ *and there exists* $\varepsilon > 0$ *such that* $h_0(t) \geq \varepsilon$ *for all* $t \in [0,\tau]$.

(III) The first derivative of $h_0(\cdot)$*, denoted by* $h_0^{(1)}(\cdot)$*, is Holder continuous with the exponent* α_0*, i.e.,* $|h_0^{(1)}(t_1) - h_0^{(1)}(t_2)| \leq M|t_1 - t_2|^{\alpha_0}$ *for some* $\alpha_0 \in (0,1]$ *and for all* $t_1, t_2 \in (0,\tau)$*, and* M *is some constant.*

Therefore, the parameter space is given by

$$\Theta = \{\theta(\cdot) \in C[0,\tau] : \theta(\cdot) \text{ satisfies (I)-(III)}\}. \tag{2.13}$$

The log-likelihood function based on the data $\boldsymbol{W} = (W_1, W_2, \ldots, W_n)$ is $l_n(\theta, \boldsymbol{W}) = \sum_{i=1}^n \log f(X_i, \Delta_i, \theta)$ where $\theta(\cdot)$ denotes the hazard function. Note that the part g of the density function p_θ is not included in the likelihood since it does not involve θ hence it does not enter into the maximum likelihood calculation. The estimate of the hazard function is obtained by maximizing $l_n(\boldsymbol{W}, \theta) = \sum_{i=1}^n l(X_i, \Delta_i, \theta)$ where $l(Y, \Delta, \theta) = \Delta \log \theta(Y) - \int_0^Y \theta(u)du$. If the maximization is taken over the infinite dimensional space Θ, it imposes great difficulty and it could lead to inconsistent estimate (Geman and Hwang, 1982). Alternatively, we can construct a subspace to approximate the parameter space Θ and carry out the maximization over this much smaller finite dimensional space so called the sieve (Grenander, 1981). Here we construct the sieve as

$$\Theta_m = \left\{ \theta_m(t) = \sum_{k=1}^m \gamma_k g_{m,k}(t) : \boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_m)^T \in [0, L_\gamma]^m \right\}, \tag{2.14}$$

where the dimension of the space, m is a function of n, say $m(n)$ and again $g_{m,k}(t) = f_\beta(t/\tau; k, m-k+1)/\tau$. We futher assume $m = m(n) = o(n^\kappa)$ for some $\kappa > 0$ that will be determined later. It follows that $\Theta_m \subseteq \Theta$ for all m. For an arbitrary element in the sieve $h_m \in \Theta_m$, we have $h_m = \sum_{k=1}^m \gamma_k g_{m,k}(t) = \sum_{k=1}^{m+1} \left\{ \frac{\gamma_k(m-k+1)}{m+1} + \frac{\gamma_{k-1}(k-1)}{m+1} \right\} g_{m+1,k}(t) \in \Theta_{m+1}$. This implies that $\Theta_m \subseteq \Theta_{m+1}$ for $m \geq 1$. As shown in the Appendix B, $d(h_m, h_0)^2$ is bounded

by $\sup_{0\le t\le\tau}|h_0(t)-h_m(t)|$, and as the true hazard rate $h_0(\cdot)$ is assumed continuous by the condition (II), so it follows from the Weierstrass theorem that there exists $\tilde{h}_m \in \Theta_m$ such that $d(\tilde{h}_m, h_0) \to 0$ as $m \to \infty$, that is, $\bigcup_{m=1}^{\infty}\Theta_m$ is dense in Θ with respect to the Hellinger metric.

Let $\hat{h}_{m,n} = \arg\max_{h\in\Theta_m} l_n(h, \boldsymbol{W})$ be the sieve maximum likelihood estimator. We know that $\hat{h}_{m,n}(\cdot)$ exists and it is unique (see *Remark 1*). The following results further imply it is also consistent and give its rate of convergence.

Theorem 1. *(Consistency) Suppose the conditions (I)-(II) hold and the sieve Θ_m is defined as in (2.14), then $\hat{h}_{m,n}(\cdot) \to_{a.s.} h_0(\cdot)$ as $m, n \to \infty$.*

Theorem 2. *(Rate of Convergence) Suppose the conditions (I)-(III) hold and the sieve Θ_m is defined as in (2.14), if $m = o(n^{\kappa})$ with $\kappa = \frac{2}{3+2\alpha_0}$, then $d(\hat{h}_{m,n}, h_0) = O_p(n^{-\frac{1+\alpha_0}{3+2\alpha_0}})$.*

The proofs of these two theorems are given in the Appendix B.

2.4 Numerical Examples

2.4.1 Simulated Data

We conducted two simulation studies to investigate the empirical performance of the proposed Bernstein polynomial based regression model and to compare it with some popular nonparametric or semiparametric models. As mentioned earlier, we focus on the estimation of the survival function, which should give clinical practitioners more information under nonproportional hazards. Therefore, we use the integrated absolute error (IAE) defined by

$$IAE = \int_0^{\tau} |\hat{S}(t) - S_0(t)|dt \approx \frac{1}{J}\sum_{j=1}^{J} |\hat{S}(t_j) - S_0(t_j)|$$

as the criterion to measure the performance of different methods studied, where $t_0 = 0, t_j = t_{j-1} + \frac{1}{J}$, t_J is a time point such that $S_0(t_J) \le .001$, $\hat{S}(\cdot)$ denotes the estimated survival function, $S_0(\cdot)$ is the true survival function, and $J = 1000$ is the number of equally-spaced time points at which we evaluate the survival functions. For both studies, we generated the random censoring

time C following an exponential distribution with rate λ_c (i.e., with the hazard $h_c(t) = \lambda_c$). We adjusted the value of λ_c in different scenarios to approximately control the censoring rate R_c to take values $\{0, 0.30, 0.50\}$, which represent completely observed, moderately censored, and severely censored data scenarios, respectively.

Binary Covariate Example: Crossing Survival Curves

The first simulation study is based on a hypothetical example of a randomized clinical trial with two treatment arms motivated by a real data set (see section 2.4.2). The event (or failure) times in both control and treatment groups were generated by log-normal distributions. In the control group the mean and standard deviation on the log-scale are -0.1 and 0.5, respectively, while these parameter are respectively set to 0 and 0.25 for the treatment group. The event times in both group were subject to independent random censoring with hazard rate λ_c adjusted to get appropriate censoring rates. The parameter values of log-normal models were deliberately chosen to allow the survival curves for two groups to cross each other during the trial. The sample size is set at $n = 100$ with equal sizes ($n_0 = n_1 = 50$) for each treatment arm. For each censoring rate, we generated $N = 1000$ Monte Carlo replicates of the data. Note that the Cox PH model is inappropriate in this setting for obvious reasons and hence not compared against our method. Instead, survival functions were estimated using our proposed Bernstein polynomial approach, the Kaplan-Meier method and the HARE method. The order of Bernstein polynomial was set at the asymptotic rate (with $\alpha_0 = 0.5$, see discussions in section 2.5 and the Appendix B) $m = [n^{0.5}] = 10$, where $[\cdot]$ denotes rounding up to the nearest integer.

The Monte Carlo pointwise means of the estimated survival curves using the three methods are shown in Figure 2.1. It is evident that the proposed model using Bernstein polynomials provides a very good fit to data generated from a model with crossing survival curves. As the Monte Carlo means approximately coincide with the true survival functions, the estimated curves obtained by the Bernstein polynomial model appear to be empirically unbiased under all three censoring scenarios. The performances of the three methods in terms of IAE are also

Table 2.1: Performances of the three methods in terms of integrated absolute error (IAE) for the simulated scenario with binary covariate. BP:Bernstein polynomial; KM:Kaplan-Meier; HARE:Hazard Regression.

	Control			Treatment		
	0%	30%	50%	0%	30%	50%
Median IAE ($\times 100$)						
BP	1.49	1.89	2.43	0.80	0.91	1.02
KM	1.79	2.26	3.05	0.87	1.10	1.29
HARE	1.96	2.27	2.58	1.08	1.39	1.59
Smallest IAE Achieved						
BP	64%	61%	54%	49%	64%	72%
KM	3%	2%	2%	25%	12%	6%
HARE	33%	37%	46%	26%	24%	22%

shown in Figure 2.2. Generally, in the control group the estimated survival functions by the Bernstein polynomial model have very similar IAE with the ones from both the Kaplan-Meier estimator and the HARE method, while the Bernstein polynomial method performs slightly better than the other two methods for the treatment group. This advantage of the Bernstein polynomial method is more clearly summarized in Table 2.1. Overall, the Bernstein polynomial method has lower median IAE than the other two methods. For each of 1000 Monte Carlo replicates, the method that has the smallest IAE is recorded. The lower panel of Table 2.1 shows the percentage out of all 1000 replicates that a particular method has the smallest IAE among the three methods being compared. As shown in Table 2.1, the Bernstein polynomial method has the smallest IAE in majority of the cases across all three censoring scenarios.

Continuous Covariate Example

For models with a continuous covariate, say Z, we studied the performance of the proposed Bernstein polynomial method when data were generated from incorrect semiparametric or parametric models. Specifically, the event time T for a given Z was generated by the conditional

log-normal nonlinear heteroscedastic model

$$\log T = \mu(Z) + \varepsilon,$$

where the mean function $\mu(Z) = \cos(\pi Z)$ and $\varepsilon|Z \sim N(0, \sigma(Z))$ with $\sigma(Z) = |Z|$. For simplicity the covariate was generated as $Z \sim Unif(0, 1)$ and the censoring time C was generated from an exponential distribution. Together with the Bernstein polynomial model and the HARE method, we fitted the data with the Cox PH model and the parametric AFT model both with the mean function $\mu(Z)$ misspecified as a linear function of Z. Also, the baseline distribution of the parametric AFT model is misspecified to be exponential distribution. Again, we set the sample size $n = 100$ and the order of Bernstein polynomial $m = [n^{0.5}] = 10$.

In Figure 2.3 it can be clearly observed that the Bernstein polynomial model provides the most accurate estimates for the conditional survival curves in this case, whereas the survival curve estimates obtained from the misspecified Cox PH model and the misspecified parametric AFT model are generally biased throughout all censoring rates. The HARE method also provides empirically unbiased estimates, but it appears that the HARE method has relatively larger variability than the Bernstein polynomial method. The results displayed in Figure 2.3 are also summarized in Figure 2.4 and Table 2.2 in terms of IAE evaluated at $Z = 0.5$. The Bernstein polynomial model outperforms the Cox PH model and the parametric AFT model by a large margin, and it also provides relatively smaller IAE than the HARE method for this particular scenario. As shown in the lower panel of Table 2.2, the Bernstein polynomial model has the smallest IAE among the four methods more than 70% of the time out of 1000 MC replicates. Thus, overall we find the performance of the Bernstein polynomial based model relatively robust especially when the dependence of the event time on other covariates is through a non-linear function.

Table 2.2: Performances of the four methods in terms of integrated absolute error (IAE) for the simulated scenario with continuous covariate. BP:Bernstein polynomial; PH: Cox Proportional Hazard Model; AFT: parametric AFT model; HARE:Hazard Regression

	Median IAE (×100)			Smallest IAE Achieved		
	0%	30%	50%	0%	30%	50%
BP	2.29	2.63	2.93	75%	76%	71%
PH	5.22	5.70	5.97	3%	6%	7%
AFT	7.83	10.13	13.51	0%	0%	0%
HARE	3.23	3.80	4.09	22%	19%	22%

2.4.2 Real Data Applications

We studied two well known datasets with potentially nonproportional hazards: the gastric cancer data and the Veterans Administration lung cancer data. The first example demonstrates the application of the proposed Bernstein polynomial model with a binary covariate, while the second one is an example of application with multiple categorical covariates and a continuous covariate.

Gastric Cancer Data

In a gastric cancer study reported by Stablein et al. (1981), $n = 90$ patients with locally advanced gastric carcinoma were randomized to two treatment groups (45 patients per group). One group only received chemotherapy while the other group received radiotherapy together with the same chemotherapy. The dataset has been studied by various authors (see e.g., Cai and Sun, 2003, and references within) and became a classic example for studying nonproportional hazards because survival curves for two treatment groups cross during the trial. As shown by the Kaplan-Meier curves in Figure 2.5, before the crossing point at approximately 1000 days the patients in the group receiving only chemotherapy had better survival rates while the benefit of combination treatment of chemotherapy and radiotherapy started to emerge at a later stage of the study. We estimated the survival functions using the model described in section 2.2.1 with the order $m = [n^{0.5}] = 10$. The results in Figure 2.5 indicate that the estimated survival curves

obtained by the proposed Bernstein polynomial model are close to the ones obtained by the Kaplan-Meier estimator. However, it is evident that the Bernstein polynomial based estimates provide much smoother survival curves which allows for a better estimation of the crossing time of the survival curves. The smooth estimated curves cross at $t = 952$ days.

Veterans Administration Data

In the Veterans Administration lung cancer study (Prentice, 1973), $n = 137$ male patients were assigned to two treatment groups: standard chemotherapy and test chemotherapy. Together with treatment assignment, 5 other baseline covariates were recorded. Following previous works (see e.g., Peng and Huang, 2007) we only used Karnofsky performance score and cell type as important covariates other than the treatment assignment. The reason for this choice can also be justified by following the initial analysis (using the Cox PH model) reported in Therneau and Grambsch (2000, chap. 6). Also, an earlier analysis based on the Cox PH model indicated that Karnofsky score and cell type are the two predictors that have significantly large test statistics for nonproportionality. Karnofsky score is a continuous variable that takes value from 0 to 100 while cell type is a categorical variable with fours levels: squamous cell, small cell, adenocarcinoma and large cell.

Since there are both categorical and continuous covariates, we first considered the stratified approach. However, given the fact that the sample size (n=137) of this study is not very large, there is a concern that there may not be sufficient data in a particular subgroup to fit the Bernstein polynomial model with continuous covariate. For example, there are only 9 patients with small cell type who received the standard chemotherapy, so Karnofsky scores of these patient might not be representative enough to span the interval $(0, 100)$. Therefore, we analyzed the data using the partially linear coefficient model described at the end of section 2.2.2 with the order $m = [n^{0.5}] = 12$. This model is a simplification of the stratified approach under the assumption that there is no interaction between Karnofsky score and cell type. At the risk of making this assumption, we were able to model the effect of the continuous covariate

Karnosky score using the entire data. The results of the estimated conditional survival contours are displayed in Figure 2.6. Generally, patients with higher Karnofsky performance scores have higher survival rates at any given time point for all groups. But such association differs across different treatment groups and more importantly across different cell types. Overall, the patients with small cell type in the test treatment group underwent the sharpest decline in survival rates, while the patients with squamous cell type receiving the test chemotherapy had the best survival profiles among all groups. The most significant treatment difference is also observed for the patients with small cell type. In this subgroup, the patients receiving the standard chemotherapy appear to have better survival rates than the ones receiving the test chemotherapy. On the contrary, patients receiving the test treatment had better survival rates than the ones receiving the standard treatment for the subgroup with squamous cell type. For the patients with adenocarcinoma or large cell type, survival contours are similar across the treatment groups.

2.5 Conclusions and Discussions

In this paper we present a nonparametric regression model for right-censored data based on Bernstein polynomials. We have demonstrated several advantages of our approach compared to other available methods in the similar context. The most remarkable feature of the proposed method is that the log-likelihood, its gradient, and the Hessian matrix all take a relatively simple form which makes it easy to compute the estimator in practice. This is due to the unique differentiability property of Bernstein polynomials. Additionally, we show that the general simple form of the log-likelihood function holds even in the presence of categorical and continuous covariates, so the proposed method can be implemented in a unified way for all these cases. Under some mild conditions, the proposed sieve maximum likelihood estimator is shown to be consistent and the corresponding rate of convergence is obtained. Through several simulation experiments with both categorical and continuous covariates, we empirically demonstrated that the proposed Bernstein polynomial model has reasonably robust performance

compared to other semiparametric models particularly when the semiparametric assumptions (e.g., PH, AFT etc.) are violated. Also, the proposed method provides similar or slightly better estimates than the HARE model in the simple situations with only binary covariates, but this advantage of the proposed method over the HARE model starts to be more prominent when the dependence of event time on covariates is relatively complex. In general, the computation algorithm involved in the proposed method was stable in the simulation studies, it reached convergence for all generated data across all scenarios studied. On the contrary, the HARE model failed to reach convergence in several occasions, especially when the censoring rate was chosen to be large. For example, in 39 out of 1000 MC replicates the HARE model did not converge for the binary covariate example when the censoring rate was 50%.

In the numerical examples throughout the paper, we used the asymptotic rate $m = [n^{0.5}]$ as the default choice for the order of Bernstein polynomial. As discussed at the end of the Appendix B, the choice of the order closely relates to the degree of smoothness or differentiability of the true hazard function. Basically, the rate $n^{2/5}$ is corresponding to the true hazard function that is twice (or more) differentiable, whereas the rate becomes $n^{2/3}$ for the true hazard function that is only differentiable to the first degree. An order that is too small will likely result in biased estimates, while a large order will introduce too much variation. So our choice of $m = [n^{0.5}]$ is a crude compromise to account for such bias-variance tradeoff. Alternatively, following Kooperberg et al. (1995) we also tried to use the Bayesian Information Criterion (BIC) to choose an optimal order. Unfortunately, the BIC approach always chose the smallest order in the given range for our method. The order m essentially plays the role of bandwidth as in kernel smoothing or number of knots as in spline smoothing. So the choice of m could also be determined by more popular data-driven approaches such as cross-validation. However, at least to the best of our efforts, we were unable to find a widely applicable cross-validation method for censored data in the regression setting. This is certainly an issue that deserves more future studies.

Lastly, although the model can contain multiple categorical covariates, only one continuous

covariate is used throughout this paper. The reason for this is that in many clinical trial settings, the number of important continuous variables is usually low, and in most cases only one continuous covariate is used. The case of multiple continuous covariates could be handled by incorporating multivariate Bernstein polynomial. For the partially linear coefficient model, additional continuous covariates can be included easily through an additive structure. However, both approaches would lead to relatively high dimensions when the number of continuous covariate is large. This is certainly a drawback of the proposed method compared to other methods such as HARE. One way around this problem would be to use a single index model to reduce the dimension of continuous covariate to a scalar and then fit the model described in this paper.

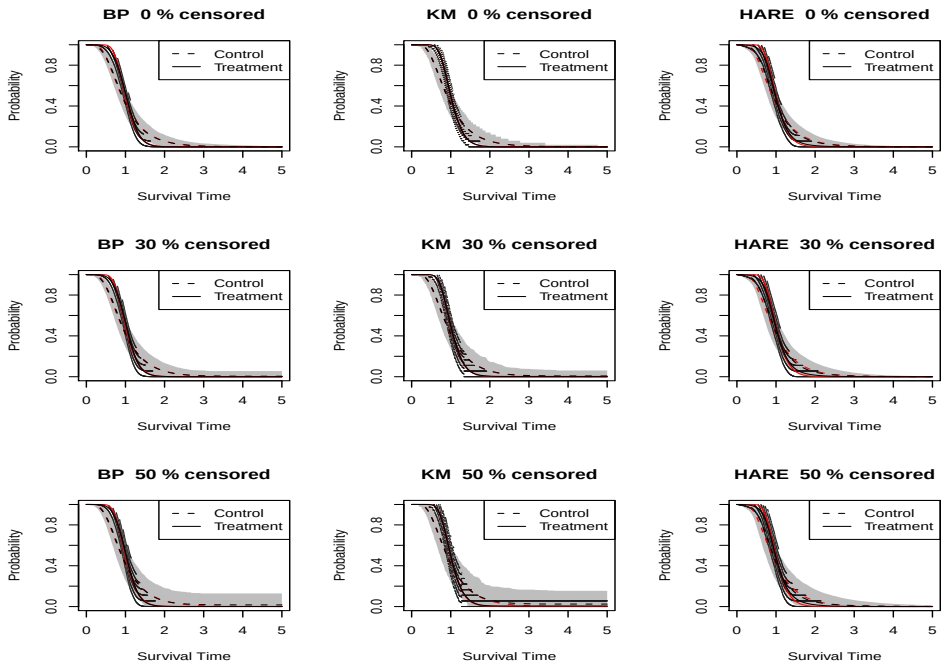


Figure 2.1: The Monte Carlo mean of the estimated survival functions. The red curves denote the true survival functions. The dotted (shaded) area denotes the area between pointwise 2.5% and 97.5% Monte Carlo percentile of survival function for the treatment (control group). BP denotes the Bernstein polynomial estimator, KM denotes the Kaplan-Meier estimator, and HARE denotes the Hazard Regression estimator.

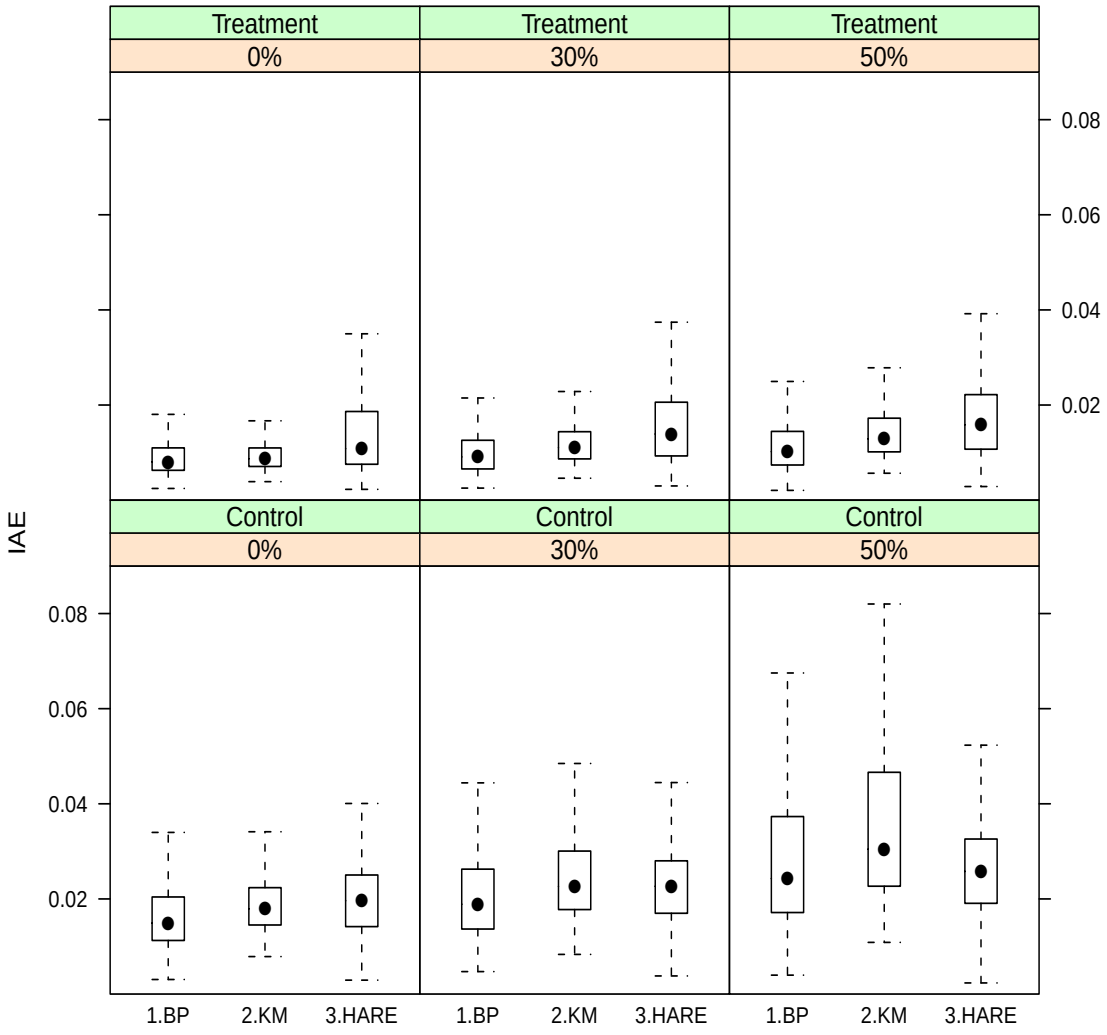


Figure 2.2: Boxplots of the integrated absolute error (IAE) for the binary covariate example based on 1000 MC replicates. The solid dot in the box represents the median IAE value. BP: the Bernstein polynomial estimator; KM: the Kaplan-Meier estimator; HARE: the Hazard Regression estimator.

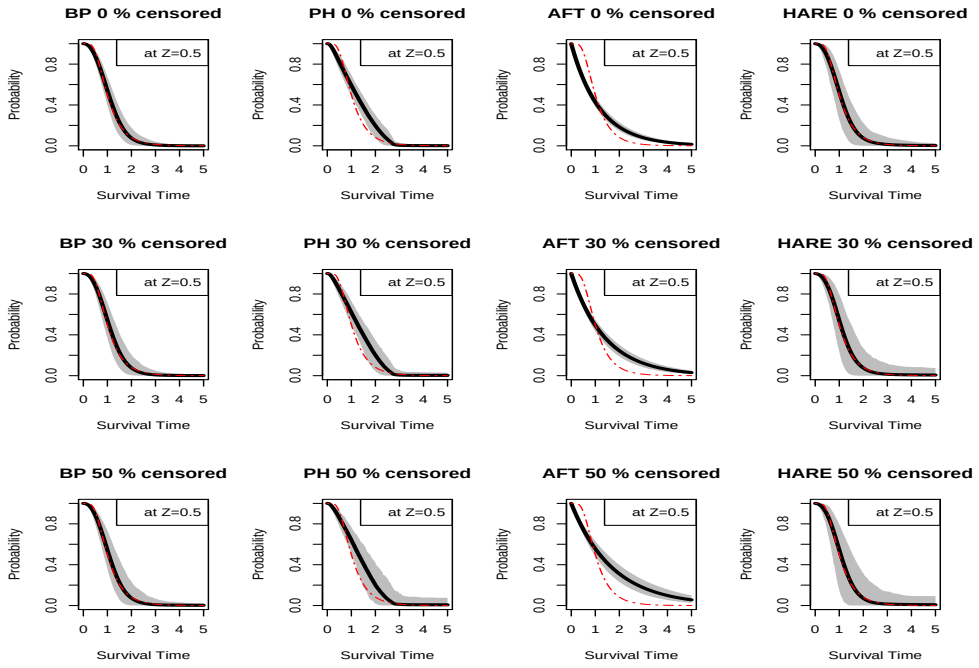


Figure 2.3: The Monte Carlo mean of the estimated conditional survival functions evaluated at $Z = \mu_z = 0.5$. The red curves denote the true survival functions. The shaded area denotes the area between pointwise 2.5% and 97.5% Monte Carlo percentile of survival function. BP denotes the Bernstein polynomial estimator, PH denotes the Cox proportional hazard estimator, paAFT denotes the parametric AFT estimator with exponential baseline, and HARE denotes the Hazard Regression estimator.

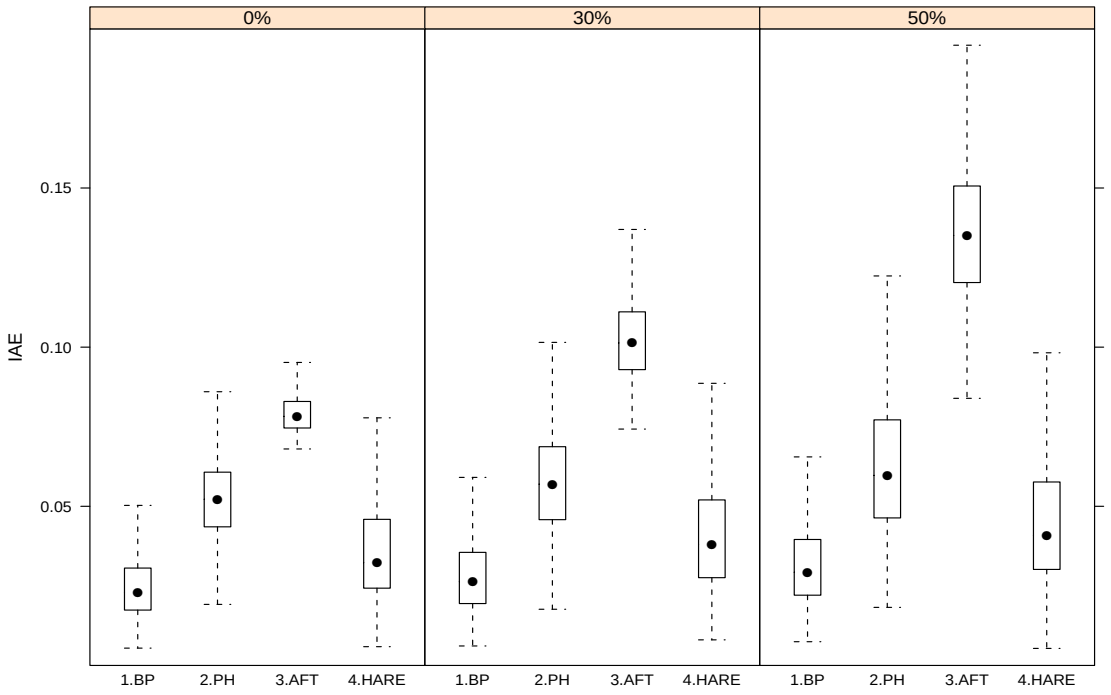


Figure 2.4: Boxplots of the integrated absolute error (IAE) of the conditional survival function evaluated at $Z = \mu_z = 0.5$. BP denotes the Bernstein polynomial estimator, PH denotes the Cox proportional hazard estimator, AFT denotes the parametric AFT estimator with exponential baseline, and HARE denotes the Hazard Regression estimator.

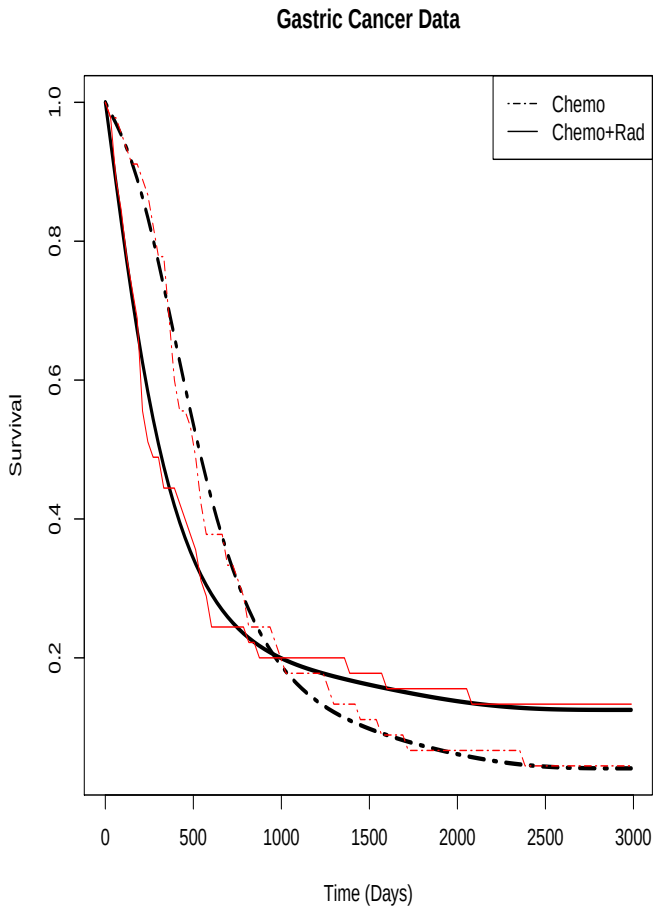


Figure 2.5: Estimated survival curves for the gastric cancer data using the Bernstein polynomial estimator (bold) and the Kaplan-Meier estimator (unbold).

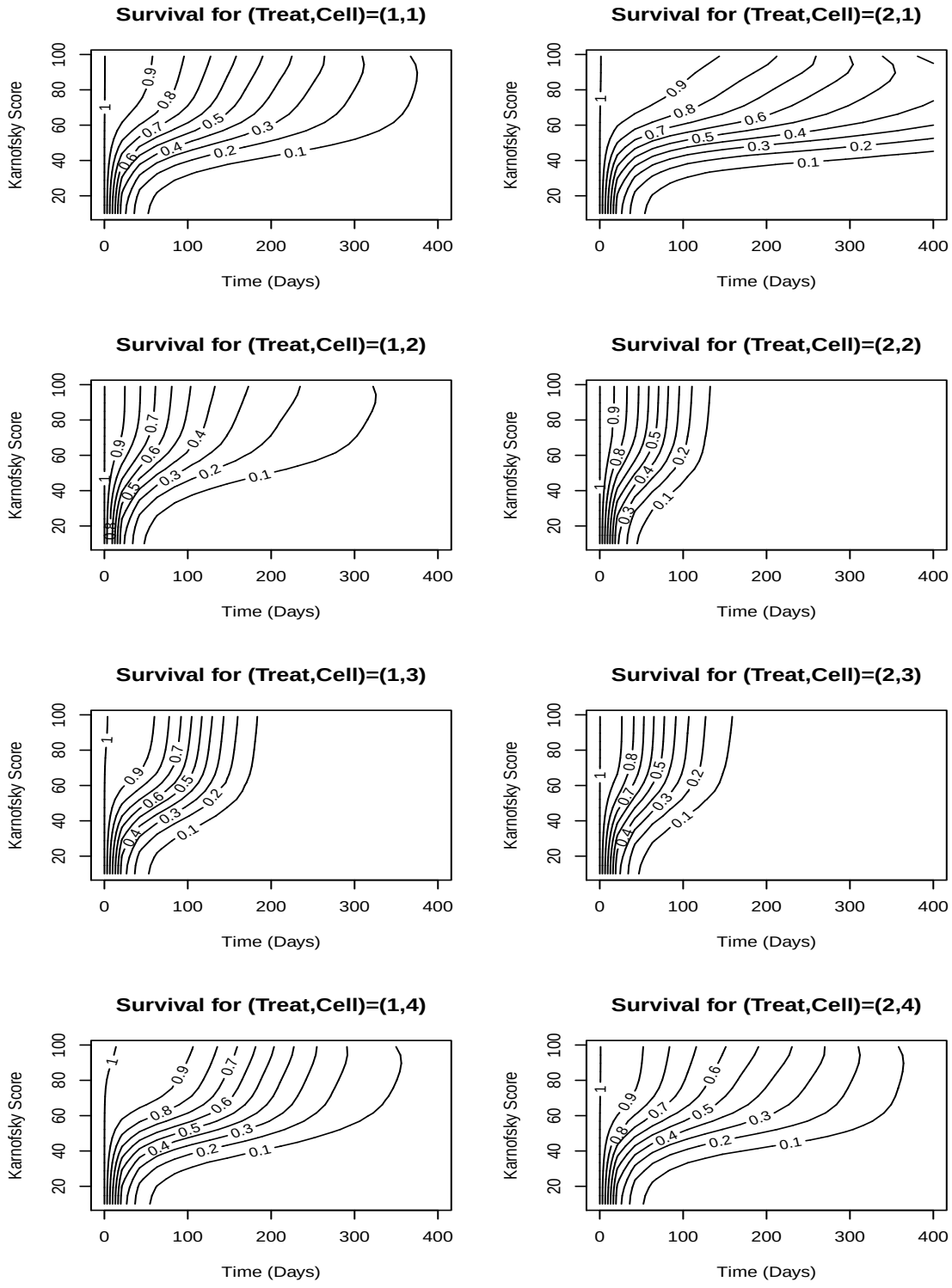


Figure 2.6: Estimated conditional survival curve contours for the Veterans Administration lung cancer data using the Bernstein polynomial estimator (the partially linear coefficient approach). "Treat" denotes treatment group: 1=standard, 2=test; "Cell" denote cell type: 1=squamous, 2=smallcell, 3=adeno, 4=large.

Chapter 3

Bayesian Nonparametric Regression Models with Varying Dimensions for Right-censored Data

3.1 Introduction

In the previous chapter a nonparametric regression model based on Bernstein polynomials is studied from a frequentist perspective. A challenging issue that still remains unresolved is that the choice of the order of the approximating polynomials is rather ambiguous. For simplicity an asymptotic rate is used. In this chapter we consider the implementation of the proposed method in a Bayesian framework. The orders of Bernstein polynomials used in the proposed model determine the dimension of the parameter space. The key feature of the Bayesian approach is its flexibility to incorporate trans-dimensional modeling by imposing a prior distribution for the order of Bernstein polynomials and update this information with data.

In the context of the proposed model with unknown number of parameters the celebrated reversible jump Markov chain Monte Carlo (MCMC) algorithm (Green, 1995) is the most rigorous approach to fit models with varying dimensions. The reversible jump MCMC sampler

can be regarded as a generalization of the Metropolis-Hastings algorithm (Matropolis et al., 1953; Hastings, 1970). Suppose the countable set of all possible models $\{\mathcal{M}_1, \mathcal{M}_2, ...\}$ is indexed by an integer k, say, the order of Bernstein polynomials in the proposed model for conditional hazard functions. The current state of the Markov chain can be denoted as (k, γ_k) where $\gamma_k \in \Gamma_k$ and Γ_k is the parameter space of the model $\mathcal{M}_k$. So the reversible jump MCMC algorithm draws samples from the whole state space $\cup_{k\in\mathfrak{K}}\{k, \Gamma_k\}$ where $\mathfrak{K}$ is the set of all possible values of k. When the model moves from the state (k, γ_k) to the state $(k', \gamma_{k'})$ with $k \neq k'$, the model space undergoes a trans-dimensional jump.

Since its introduction the reversible jump MCMC approach has been studied in many areas especially in finite mixture models (Richardson and Green, 1997) and variable selection problems (Sillanpana and Arjas, 1998). Also, there has been increasing uses of the method in genetics and computational biology (see Waagepetersen and Sorensenz, 2001). Another important application area of the reversible jump MCMC is function estimation. In the most relevant context, Chang et al. (2005) implemented the reversible jump MCMC for hazard estimation without covariates based on a model using Bernstein polynomials. In this chapter, we apply the reversible jump MCMC sampler to the regression models with categorical or continuous covariates. The rest of the chapter proceeds as follows. In section 3.2, we describe the Bayesian regression models for categorical and continuous covariates. We specify the prior distributions and describe the details of posterior sampling scheme. We first apply the proposed Bayesian hazard regression model to a simulated dataset in section 3.3. Further, the proposed method is demonstrated using real data from two cancer studies. Finally, we conclude with some discussion in section 3.4.

3.2 Bayesian Nonparametric Conditional Hazard Models

Recall T_i denotes the time to certain event of interest for the subject i and Z_i denotes a vector of covariates for $i = 1, 2, ..., n$. The event time T_i is subject to random right-censoring C_i and hence for every subject we observe (X_i, Δ_i), where $X_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$.

Further, we assume for each i, T_i is conditionally independent of C_i given Z_i for $i = 1, 2, ..., n$. For any $t \geq 0$, the conditional cumulative hazard function is given by $H(t|Z) = -\log S(t|Z)$ and hazard function $h(t|Z) = \dot{H}(t|Z)$, where $S(t|Z) = Pr(T_i > t|Z_i = z)$ is the survival function. Throughout the chapter we assume that the triplet (T_i, C_i, Z_i) are independent and identically distributed. Further, we assume that there exists a $\tau(Z) < \infty$ such that $\tau(Z) = \inf\{t : S(t|Z) = 0\}$.

As shown in the previous chapter, the expression for the hazard function and the cumulative hazard function can be written in a unified form for the cases with no covariates, categorical covariates, and continuous covariates as

$$h_k(t, \boldsymbol{\gamma}|Z_i) = U_i^T \boldsymbol{\gamma} \text{ and } H_k(t, \boldsymbol{\gamma}|Z_i) = V_i^T \boldsymbol{\gamma},$$

where k denotes total number of parameters, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_k)^T$, and U_i and V_i are some vectors depending only on the data.

1. When the covarate Z_i is a categorical variable with levels : $1, 2, 3, ..., J$, we have $k = mJ$ and

$$U_i = \begin{bmatrix} I(Z_i = 1)\boldsymbol{g}_m(X_i) \\ I(Z_i = 2)\boldsymbol{g}_m(X_i) \\ \vdots \\ I(Z_i = J)\boldsymbol{g}_m(X_i) \end{bmatrix} \text{ and } V_i = \begin{bmatrix} I(Z_i = 1)\boldsymbol{G}_m(X_i) \\ I(Z_i = 2)\boldsymbol{G}_m(X_i) \\ \vdots \\ I(Z_i = J)\boldsymbol{G}_m(X_i) \end{bmatrix},$$

   where $\boldsymbol{g}_m(X_i) = (g_{m1}(X_i), g_{m2}(X_i), ..., g_{ml}(X_i), ..., g_{mm}(X_i))^T$, $g_{ml}(X_i) = f_\beta(X_i/\tau; l, m - l + 1)/\tau$, $\boldsymbol{G}_m(X_i) = (G_{m1}(X_i), G_{m2}(X_i), ..., G_{ml}(X_i), ..., G_{mm}(X_i))^T$, $G_{ml}(X_i) = F_\beta(X_i/\tau; l, m - l + 1)$, and $f_\beta(\cdot; l, m - l + 1)$ and $F_\beta(\cdot; l, m - l + 1)$ are respectively the probability density function (PDF) and the cumulative distribution function (CDF) of the Beta distribution with shape parameters l and $m - l + 1$.

2. When the covariate Z_i is a continuous variable (assuming that it has already been mapped

to the unit interval $[0, 1]$, see chapter 2 section 2.2.4 for details), we have $k = m\tilde{m}$

$$U_i = vec[\tilde{\boldsymbol{g}}_{\tilde{m}}(Z_i)\boldsymbol{g}_m^T(X_i)] \text{ and } V_i = vec[\tilde{\boldsymbol{g}}_{\tilde{m}}(Z_i)\boldsymbol{G}_m^T(X_i)]$$

where $\tilde{\boldsymbol{g}}_{\tilde{m}}(Z_i) = (\tilde{g}_{\tilde{m}1}(Z_i), \tilde{g}_{\tilde{m}2}(Z_i), ..., \tilde{g}_{\tilde{m}j}(Z_i), ..., g_{\tilde{m}\tilde{m}}(Z_i))^T$, $\tilde{g}_{\tilde{m}j}(Z_i) = f_\beta(Z_i; j, \tilde{m} - j + 1)/\tilde{m}$, and $vec[\cdot]$ denotes the vectorization by column operator applied to a matrix.

3. When both a categorical covariate $Z1_i$ and a continuous covariate $Z2_i$ are present, $k = (J + \tilde{m})m$ and we assume an additive structure

$$U_i = [U1_i^T, U2_i^T]^T \text{ and } V_i = [V1_i^T, V2_i^T]^T,$$

with

$$U1_i = \begin{bmatrix} I(Z1_i = 1)\boldsymbol{g}_m(X_i) \\ I(Z1_i = 2)\boldsymbol{g}_m(X_i) \\ \vdots \\ I(Z1_i = J)\boldsymbol{g}_m(X_i) \end{bmatrix} \text{ and } V1_i = \begin{bmatrix} I(Z1_i = 1)\boldsymbol{G}_m(X_i) \\ I(Z1_i = 2)\boldsymbol{G}_m(X_i) \\ \vdots \\ I(Z1_i = J)\boldsymbol{G}_m(X_i) \end{bmatrix},$$

and

$$U2_i = vec[\tilde{\boldsymbol{g}}_{\tilde{m}}(Z2_i)\boldsymbol{g}_m^T(X_i)] \text{ and } V2_i = vec[\tilde{\boldsymbol{g}}_{\tilde{m}}(Z2_i)\boldsymbol{G}_m^T(X_i)].$$

Conveniently, the log-likelihood function in all cases can be written in a general form as $l(\boldsymbol{\gamma}|\boldsymbol{X}) = \sum_{i=1}^n \{\Delta_i \log(U_i^T\boldsymbol{\gamma}) - V_i^T\boldsymbol{\gamma}\}$.

3.2.1 Prior Specification

For a Bayesian approach, the prior distributions of the parameters are crucial. Now we specify the prior distributions for the vector of parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_k)$, where k denotes the total number of parameters. Recall the physical interpretation of γ_j is the increment in the cumulative hazard function from the knot $j - 1$ to the knot j. Since the cumulative hazard function is a non-decreasing function, the restriction on the parameters is that $\gamma_j \geq 0$ for

$j = 1, 2, ..., k$. When $\gamma_j > 0$, a log-normal distribution is assumed as the prior distribution. It is also important to capture boundary values $\gamma_j = 0$ for $j = 1, 2, ..., k$. Therefore, we consider a two-component mixture prior for γ_j with a positive probability mass at 0. The prior density function can expressed as

$$\pi(\gamma_j) = p_0 \boldsymbol{I}(\gamma_j = 0) + (1 - p_0)\boldsymbol{I}(\gamma_j > 0) f_{LN}(\gamma_j; a, b) \text{ for } j = 1, 2, ..., k$$

where $p_0 \in (0, 1)$ denotes the weight in the mixture prior, $\boldsymbol{I}(\cdot)$ is the indicator function, and $f_{LN}(\gamma; a, b)$ is the log-normal density function with mean a and standard deviation b on the log scale. This prior specification shares a great similarity with the stochastic search variable selection (SSVS) approach considered by George and McCulloch (1993) and Geweke (1996). So the key advantage of this prior specification is to allow for sparsity when we overestimate the number of unique nodes in the Bernstein approximation. For a variable follows a mixture distribution, it is oftentimes convenient in the sampling process if we treat the mixture random variable as a product of two independent variables (see Kuo and Mallick, 1998). Accordingly, we can regard the parameters of interest as

$$\gamma_j = \delta_j \alpha_j,$$

where δ_j has a Bernoulli prior distribution with the proportion p_0 (denoted as $\pi_\delta(\delta_j)$) and the prior for α_j is a log-normal distribution with mean a and standard deviation b on the log scale (denoted as $\pi_\alpha(\alpha_j|a, b)$). For simplicity, we assume α_j an δ_j are i.i.d and mutually independent for $j = 1, 2, ..., k$.

3.2.2 Posterior Sampling Scheme

In order to make posterior inference, we need to draw samples from the posterior distribution expressed as

$$p(\boldsymbol{\gamma}|\boldsymbol{X}) \propto \exp[l(\boldsymbol{\gamma}|\boldsymbol{X})]\pi_\gamma(\boldsymbol{\gamma}), \tag{3.1}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_p)^T$ with $\gamma_j = \delta_j \alpha_j$ and $\pi_\gamma(\boldsymbol{\gamma}) = \prod_{j=1}^{k} \pi(\gamma_j)$.

Fixed-order Scenario

We first describe the scenario when the order k is assumed fixed. Similar to that of the SSVS approach, the full conditional distribution can derived for the indicator variables δ_j, $j = 1, 2, ..., k$ as

$$p(\delta_j = 1|\text{rest}) = \frac{(1-p_0)\exp[l_{\delta_j=1}(\boldsymbol{\gamma}|\boldsymbol{X})]}{(1-p_0)\exp[l_{\delta_j=1}(\boldsymbol{\gamma}|\boldsymbol{X})] + p_0 \exp[l_{\delta_j=0}(\boldsymbol{\gamma}|\boldsymbol{X})]},$$

where $l_{\delta_j=1}(\boldsymbol{\gamma}|\boldsymbol{X})$ and $l_{\delta_j=0}(\boldsymbol{\gamma}|\boldsymbol{X})$ denote the values of the log-likelihood function evaluated by setting $\delta_j = 1$ ($\gamma_j = \delta_j \alpha_j = \alpha_j$) and $\delta_j = 0$ ($\gamma_j = \delta_j \alpha_j = 0$), respectively. Therefore, the posterior simulation for δ_j, $j = 1, 2, ..., k$ is straight forward via Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) as the full conditional take a simple form that we can directly draw samples.

The full conditional for α_j, $j = 1, 2, ..., k$, can be written as

$$p(\alpha_j|\text{rest}) \propto \begin{cases} \exp[l_{\delta_j=1}(\boldsymbol{\gamma}|\boldsymbol{X})]\pi_\alpha(\alpha_j|a,b) & \text{if } \delta_j = 1 \\ \pi_\alpha(\alpha_j|a,b) & \text{if } \delta_j = 0 \end{cases},$$

where α_j's prior density $\pi_\alpha(\alpha_j|a,b)$ is lognormal with mean a and standard deviation b on the log scale. When $\delta_j = 0$ one can easily draw samples from the full conditional, however, the full conditional of α_j can not be directly sampled when $\delta_j = 1$. For this situation we apply Metroplis-Hastings algorithm to generate a sample from a proposal distribution and accept it with certain probability. Specifically, the proposal distribution $q(\cdot; \log(\alpha_j^*), \sigma_0)$ is a log-normal distribution with mean $\log(\alpha_j^*)$ and standard deviation σ_0 on the log scale, where α_j^* is the state of α_j in the previous iteration. This proposal distribution is similar to that of the random walk proposals except that we use log-normal distribution because α_j is always positive. After

a proposal is drawn, we accept it with a probability

$$\min\left(\frac{p(\alpha_j|\text{rest})}{p(\alpha_j^*|\text{rest})}\frac{q(\alpha_j^*,\log(\alpha_j),\sigma_0)}{q(\alpha_j,\log(\alpha_j^*),\sigma_0)},1\right)$$

The sampling scheme for the fixed-order scenario can be summarized as following.

Algorithm 1. *Suppose* $\alpha_j^{(t)}$ *and* $\delta_j^{(t)}$ *are the current states at the t-th iteration for* $j = 1, 2, .., k$ *and we set the initial values to be* $\alpha_j^{(0)}$ *and* $\delta_j^{(0)}$ *for* $j = 1, 2, .., k$.

1. *Generate* $\delta_j^{(t)}$ *with the full conditional* $p(\delta_j^{(t)} = 1|rest)$ *for* $j = 1, 2, ..., k$
2. - *If* $\delta_j^{(t)} = 0$ *then generate* $\alpha_j^{(t)}$ *with the prior* $\pi_\alpha(\alpha_j^{(t)}; a, b)$.
   - *If* $\delta_j^{(t)} = 1$ *then generate a proposal* $\alpha_j^{(t)}$ *with* $q(\alpha_j^{(t)}; \log(\alpha_j^{(t-1)}))$ *and accept it as the current state with probability*

$$\min\left(\frac{p(\alpha_j^{(t)}|rest)}{p(\alpha_j^{(t-1)}|rest)}\frac{q(\alpha_j^{(t-1)},\log(\alpha_j^{(t)}),\sigma_0)}{q(\alpha_j^{(t)},\log(\alpha_j^{(t-1)}),\sigma_0)},1\right)$$

   otherwise set $\alpha_j^{(t)} = \alpha_j^{(t-1)}$

   - *Repeat this for all* $j = 1, 2, ..., k$

The above algorithm can be further optimized by sampling the continuous parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_p)^T$ together. One advantage of blocking is to reduce autocorrelation and hence speed up convergence. For the vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k)^T$ the full conditional can be written as

$$p(\boldsymbol{\alpha}|\text{rest}) \propto \pi(\boldsymbol{\alpha}) \exp(l(\boldsymbol{\alpha} \cdot \boldsymbol{\delta}|\boldsymbol{X})],$$

where $\pi(\boldsymbol{\alpha}) = \prod_{j=1}^{k} \pi_\alpha(\alpha_j)$ and $\boldsymbol{\alpha} \cdot \boldsymbol{\delta}$ denotes the Hadamard or element-wise product of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$. With this full conditional, the proposal can be drawn as a vector from the multi-variate log-normal distribution $q_{vec}(\cdot; \log(\boldsymbol{\alpha}^*), \sigma_0^2\boldsymbol{I})$ where $\boldsymbol{I}$ denote $k \times k$ identity matrix. Therefore, the algorithm 1 can be modified to

Algorithm 2. *Suppose* $\boldsymbol{\alpha}^{(t)}$ *and* $\boldsymbol{\delta}^{(t)}$ *are the current states at the* t*-th iteration, and we set the initial values to be* $\boldsymbol{\alpha}^{(0)}$ *and* $\boldsymbol{\delta}^{(0)}$*.*

1. *Generate* $\delta_j^{(t)}$ *with the full conditional* $p(\delta_j^{(t)} = 1|rest)$ *for* $j = 1, 2, ..., k$
2. *Generate a proposal vector* $\boldsymbol{\alpha}^{(t)}$ *with* $q_{vec}(\boldsymbol{\alpha}^{(t)}; \log(\boldsymbol{\alpha}^{(t-1)}), \sigma_0^2 \boldsymbol{I})$ *and accept it as the current state with probability*

$$\min\left(\frac{p(\boldsymbol{\alpha}^{(t)}|rest)}{p(\boldsymbol{\alpha}^{(t-1)}|rest)}\frac{q_{vec}(\boldsymbol{\alpha}^{(t-1)}, \log(\boldsymbol{\alpha}^{(t)}), \sigma_0^2\boldsymbol{I})}{q_{vec}(\boldsymbol{\alpha}^{(t)}, \log(\boldsymbol{\alpha}^{(t-1)}), \sigma_0^2\boldsymbol{I})}, 1\right),$$

otherwise set $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(t-1)}$*.*

Reversible Jump MCMC

When the order k is allowed to change, the proposed nonparametric regression model becomes a variable dimension model. Suppose $\boldsymbol{\gamma}_k$ denotes the vector of the all parameters under the k-dimensional model, the joint distribution of k, $\boldsymbol{\gamma}_k$ and the data $\boldsymbol{X}$ can be expressed as

$$p(k, \boldsymbol{\gamma}_k, \boldsymbol{X}) = p(k)p(\boldsymbol{\gamma}_k|k)p(\boldsymbol{X}|\boldsymbol{\gamma}_k, k),$$

where $p(k)$ denotes the prior probability of k, $p(\boldsymbol{\gamma}_k|k)$ denotes the prior probability of $\boldsymbol{\gamma}_k$ given k, and $p(\boldsymbol{X}|\boldsymbol{\gamma}_k, k)$ is the likelihood (Green, 1995). Since k is now a component of unknown parameters in the model, the posterior distribution becomes

$$p(k, \boldsymbol{\gamma}_k|\boldsymbol{X}) \propto p(k)p(\boldsymbol{\gamma}_k|k)p(\boldsymbol{X}|\boldsymbol{\gamma}_k, k).$$

The reversible jump MCMC approach proposed by Green (1995) provided a solution to sample from a state space with varying dimensions. This whole state space can be denoted as $\cup_{k\in\mathfrak{K}}\{k, \Gamma_k\}$ where $\mathfrak{K}$ is the set of all possible values of k and Γ_k is the parameter space for the k-dimensional model. The key feature of the reversible jump MCMC is a technique called dimension matching. Basically, when a transition is proposed from a state $(k, \boldsymbol{\gamma}_k)$ to another

state $(k', \boldsymbol{\gamma}_{k'})$ with different dimensions (i.e. $k \neq k'$), Green (1995) proposed a bijection $T(\cdot, \cdot)$ with data augmentation u and v to map $(k, \boldsymbol{\gamma}_k)$ to $(k', \boldsymbol{\gamma}_{k'})$, that is, $(\gamma_{k'}, v) = T_{k \to k'}(\gamma_k, u)$ with $\dim(u) + k = \dim(v) + k'$, where u and v are random vectors generated by densities $p_u(\cdot)$ and $p_v(\cdot)$, respectively. For the theoretical justification and details about dimension matching and reversible jump MCMC readers can refer to the seminal article by Green (1995). Similar to the Metropolis-Hastings sampler, after a move from $(k, \boldsymbol{\gamma}_k)$ to $(k', \boldsymbol{\gamma}_{k'})$ is proposed we accept it as the current state with probability

$$\min\left(\frac{p(k)p(\boldsymbol{\gamma}_k|k)p(\boldsymbol{X}|\boldsymbol{\gamma}_k, k)}{p(k')p(\boldsymbol{\gamma}_{k'}|k')p(\boldsymbol{X}|\boldsymbol{\gamma}_{k'}, k')}\frac{p_{k' \to k}p_v(v)}{p_{k \to k'}p_u(u)}\left|\frac{\partial T_{k \to k'}(\gamma_k, u)}{\partial(\gamma_k, u)}\right|, 1\right),$$

where $p_{k \to k'}$ denotes the probability of the jump from k-dimension to k'-dimension while $p_{k' \to k}$ is the probability jumping from k'-dimension to k-dimension.

The most important step of the reversible jump sampler in practice is to propose a most suitable way of dimension matching given the context of the problem. In this chapter, we use an approach known as the "birth and death" process to explore the state spaces with different dimensions. Generally, if $k' > k$ when we propose a move from $(k, \boldsymbol{\gamma}_k)$ to $(k', \boldsymbol{\gamma}_{k'})$ then the "birth" process is applied by generating a random vector u with $\dim(u) = k' - k$. For the data generating probability of u we use the prior distribution of the corresponding parameters. When $k' < k$ the "death" process is applied by eliminating $k - k'$ components of the vector of parameters. One advantage of using a "birth-and-death" process is that the Jacobian in the acceptance probability becomes 1. The acceptance probability of a proposed trans-dimensional move depends more on likelihood ratio instead of the behavior of the synthesized data. Also, in this chapter we assume a discrete uniform distribution for the order k such that $\frac{p(k)}{p(k')}$ always equals to 1. In order to further simply the expression for the acceptance rate we set $p_{k \to k'} = p_{k' \to k}$, that is, we assign equal probabilities to the move from k to k' and the move k' to k.

Now we describe the sampling scheme for the proposed model with a continuous covariate. In this case, the dimension of the proposed nonparametric regression model $k = m\tilde{m}$, where

m denotes the order of the Bernstein polynomials for the time variable and $\tilde{m}$ denotes the order of the Bernstein polynomials for the continuous covariate. The vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ can be re-expressed in matrix forms

$$\boldsymbol{\alpha}_M = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,\tilde{m}} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,\tilde{m}} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m,1} & \alpha_{m,2} & \cdots & \alpha_{m,\tilde{m}} \end{pmatrix} \text{ and } \boldsymbol{\delta}_M = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,\tilde{m}} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,\tilde{m}} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{m,1} & \delta_{m,2} & \cdots & \delta_{m,\tilde{m}} \end{pmatrix}.$$

A change in m results in a change in the number of rows while a change in $\tilde{m}$ means a change in the number of columns. As suggested by Richardson and Green (1997), the trans-dimensional moves should be restricted to the models that are not too far apart in the state space. Similar to the Metropolis-Hastings sampler a proposal move the reversible jump MCMC that is too far away from the previous state results in low acceptance rate and hence inefficient exploration of the state space during sampling. Therefore, the trans-dimensional moves in the chapter only can be one of the following: a row increase ($m \to m+1$), a row decrease ($m \to m-1$), a column increase ($\tilde{m} \to \tilde{m}+1$), and a column decrease ($\tilde{m} \to \tilde{m}-1$). When a move to lower dimension is proposed, data augmentation is not needed. However, when a move to higher dimension is proposed we need to generate a set of new values for the additional row or column of $\boldsymbol{\alpha}_M$ and $\boldsymbol{\delta}_M$ using the prior probabilities $\pi_\alpha(\cdot; a, b)$ and $\pi_\delta(\cdot; p_0)$, respectively. The above processes can be summarized as

Algorithm 3. *Suppose $\boldsymbol{\gamma}^{(t)}$, $\boldsymbol{\alpha}^{(t)}$ (with matrix representation $\boldsymbol{\alpha}_M^{(t)}$), and $\boldsymbol{\delta}^{(t)}$ (with matrix representation $\boldsymbol{\delta}_M^{(t)}$) are the current states at the t-th iteration, and we set the initial values to be $\boldsymbol{\alpha}^{(t)}$, $\boldsymbol{\delta}^{(t)}$, and $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\alpha}^{(t)} \cdot \boldsymbol{\delta}^{(t)}$.*

1. *select one from the following moves with equal probabilities*
   - *move 0: parameter-updating ($m \to m; \tilde{m} \to \tilde{m}$)*
   - *move 1: add a row (birth: $m \to m+1$)*

- *move 2: remove a row (death: $m \to m-1$)*
- *move 3: add a column (birth: $\tilde{m} \to \tilde{m}+1$)*
- *move 4: remove a column (death: $\tilde{m} \to \tilde{m}-1$)*

2. *if the move 0 is selected, update parameters using algorithm 1 or algorithm 2*

3. *if the move 1 is selected*

   - *generate $\boldsymbol{u}_1 = (u_{11}, u_{12}, \ldots, u_{1\tilde{m}})^t$ with $u_{1j} \sim \pi_\alpha(\cdot; a, b)$ and $\boldsymbol{u}_2 = (u_{21}, u_{22}, \ldots, u_{2\tilde{m}})^t$ with $u_{2j} \sim \pi_\delta(\cdot; p_0)$*
   - *construct $\boldsymbol{\alpha}_M^{(t)}$ by adding the row $\boldsymbol{u}_1$ to $\boldsymbol{\alpha}_M^{(t-1)}$ as the last row, construct $\boldsymbol{\delta}_M^{(t)}$ by adding the row $\boldsymbol{u}_2$ to $\boldsymbol{\delta}_M^{(t-1)}$ as the last row and construct $\boldsymbol{\gamma^{(t)}} = vec[\boldsymbol{\alpha}_M^{(t)} \cdot \boldsymbol{\delta}_M^{(t)}]$ (where $\boldsymbol{\alpha}_M^{(t-1)} \cdot \boldsymbol{\delta}_M^{(t-1)}$ means the Hadamard or element-wise product of $\boldsymbol{\alpha}_M^{(t-1)}$ and $\boldsymbol{\delta}_M^{(t-1)}$); accept $\boldsymbol{\alpha}_M^{(t)}$, $\boldsymbol{\delta}_M^{(t)}$, $\boldsymbol{\gamma^{(t)}}$ as the current states with probability*

$$\min\left(\frac{p(\boldsymbol{\gamma}^{(t)}|\boldsymbol{X})}{p(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{X})}\frac{1}{\prod_{j=1}^{\tilde{m}}[\pi_\alpha(u_{1j}; a, b)p_0^{u_{2j}}(1-p_0)^{(1-u_{2j})}]}, 1\right),$$

   otherwise $\boldsymbol{\alpha}_M^{(t)} = \boldsymbol{\alpha}_M^{(t-1)}$, $\boldsymbol{\delta}_M^{(t)} = \boldsymbol{\delta}_M^{(t-1)}$ and $\boldsymbol{\gamma^{(t)}} = \boldsymbol{\gamma^{(t-1)}}$.

4. *if the move 2 is selected*

   - *generate $\boldsymbol{u}_1 = (u_{11}, u_{12}, \ldots, u_{1\tilde{m}})^t$ with $u_{1j} \sim \pi_\alpha(\cdot; a, b)$ and $\boldsymbol{u}_2 = (u_{21}, u_{22}, \ldots, u_{2\tilde{m}})^t$ with $u_{2j} \sim \pi_\delta(\cdot; p_0)$*
   - *construct $\boldsymbol{\alpha}_M^{(t)}$ by removing the last row of $\boldsymbol{\alpha}_M^{(t-1)}$, construct $\boldsymbol{\delta}_M^{(t)}$ by removing the last row of $\boldsymbol{\delta}_M^{(t-1)}$, and construct $\boldsymbol{\gamma^{(t)}} = vec[\boldsymbol{\alpha}_M^{(t)} \cdot \boldsymbol{\delta}_M^{(t)}]$; accept $\boldsymbol{\alpha}_M^{(t)}$, $\boldsymbol{\delta}_M^{(t)}$, $\boldsymbol{\gamma^{(t)}}$ as the current states with probability*

$$\min\left(\frac{p(\boldsymbol{\gamma}^{(t)}|\boldsymbol{X})}{p(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{X})}\prod_{j=1}^{\tilde{m}}[\pi_\alpha(u_{1j}; a, b)p_0^{u_{2j}}(1-p_0)^{(1-u_{2j})}], 1\right),$$

   otherwise $\boldsymbol{\alpha}_M^{(t)} = \boldsymbol{\alpha}_M^{(t-1)}$, $\boldsymbol{\delta}_M^{(t)} = \boldsymbol{\delta}_M^{(t-1)}$ and $\boldsymbol{\gamma^{(t)}} = \boldsymbol{\gamma^{(t-1)}}$.

5. if the move 3 is selected

- *generate* $\boldsymbol{u}_1 = (u_{11}, u_{12}, \ldots, u_{1m})^t$ *with* $u_{1j} \sim \pi_\alpha(\cdot; a, b)$ *and* $\boldsymbol{u}_2 = (u_{21}, u_{22}, \ldots, u_{2m})^t$ *with* $u_{2j} \sim \pi_\delta(\cdot; p_0)$
- *construct* $\boldsymbol{\alpha}_M^{(t)}$ *by adding the column* $\boldsymbol{u}_1$ *to* $\boldsymbol{\alpha}_M^{(t-1)}$ *as the last column, construct* $\boldsymbol{\delta}_M^{(t)}$ *by adding the column* $\boldsymbol{u}_2$ *to* $\boldsymbol{\delta}_M^{(t-1)}$ *as the last column and construct* $\boldsymbol{\gamma^{(t)}} = vec[\boldsymbol{\alpha}_M^{(t)} \cdot \boldsymbol{\delta}_M^{(t)}]$ *(where* $\boldsymbol{\alpha}_M^{(t-1)} \cdot \boldsymbol{\delta}_M^{(t-1)}$ *means the Hadamard or element-wise product of* $\boldsymbol{\alpha}_M^{(t-1)}$ *and* $\boldsymbol{\delta}_M^{(t-1)}$*); accept* $\boldsymbol{\alpha}_M^{(t)}$, $\boldsymbol{\delta}_M^{(t)}$, $\boldsymbol{\gamma^{(t)}}$ *as the current states with probability*

$$\min\left(\frac{p(\boldsymbol{\gamma}^{(t)}|\boldsymbol{X})}{p(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{X})}\frac{1}{\prod_{j=1}^{m}[\pi_\alpha(u_{1j}; a, b)p_0^{u_{2j}}(1-p_0)^{(1-u_{2j})}]}, 1\right),$$

otherwise $\boldsymbol{\alpha}_M^{(t)} = \boldsymbol{\alpha}_M^{(t-1)}$, $\boldsymbol{\delta}_M^{(t)} = \boldsymbol{\delta}_M^{(t-1)}$ *and* $\boldsymbol{\gamma^{(t)}} = \boldsymbol{\gamma^{(t-1)}}$.

6. if the move 4 is selected

- *generate* $\boldsymbol{u}_1 = (u_{11}, u_{12}, \ldots, u_{1m})^t$ *with* $u_{1j} \sim \pi_\alpha(\cdot; a, b)$ *and* $\boldsymbol{u}_2 = (u_{21}, u_{22}, \ldots, u_{2m})^t$ *with* $u_{2j} \sim \pi_\delta(\cdot; p_0)$
- *construct* $\boldsymbol{\alpha}_M^{(t)}$ *by removing the last column of* $\boldsymbol{\alpha}_M^{(t-1)}$*, construct* $\boldsymbol{\delta}_M^{(t)}$ *by removing the last column of* $\boldsymbol{\delta}_M^{(t-1)}$*, and construct* $\boldsymbol{\gamma^{(t)}} = vec[\boldsymbol{\alpha}_M^{(t)} \cdot \boldsymbol{\delta}_M^{(t)}]$ *; accept* $\boldsymbol{\alpha}_M^{(t)}$, $\boldsymbol{\delta}_M^{(t)}$, $\boldsymbol{\gamma^{(t)}}$ *as the current states with probability*

$$\min\left(\frac{p(\boldsymbol{\gamma}^{(t)}|\boldsymbol{X})}{p(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{X})}\prod_{j=1}^{m}[\pi_\alpha(u_{1j}; a, b)p_0^{u_{2j}}(1-p_0)^{(1-u_{2j})}], 1\right),$$

otherwise $\boldsymbol{\alpha}_M^{(t)} = \boldsymbol{\alpha}_M^{(t-1)}$, $\boldsymbol{\delta}_M^{(t)} = \boldsymbol{\delta}_M^{(t-1)}$ *and* $\boldsymbol{\gamma^{(t)}} = \boldsymbol{\gamma^{(t-1)}}$.

Note that in the above algorithm we include a move in which the dimension of the model remains the same as its previous state (Move 0: parameter-updating). As pointed out by Marin and Robert (2007), although such a move is not necessary in the reversible jump sampler, adding fixed-dimensional moves can greatly increase the sampling efficiency. The procedure for cases with categorical covaraites is a special case of the above algorithm. For these cases the column

number of $\boldsymbol{\alpha}_M$ and $\boldsymbol{\delta}_M$ is fixed at the number of categories J. So in the dimension jump process, only moves adding or removing rows are allowed. When there are both a categorical covariate and a continuous present in the regression model the algorithm 3 can be readily extended to incorporate the dimension changes. Basically, we partition the parameters to the ones associated with the categorical covariate and the ones associated with covariate and construct two matrices for α's and two matrices for δ's. So when a jump is proposed for m or $\tilde{m}$ we update these matrices accordingly.

3.3 Numerical Examples

Now we demonstrate the Bayesian implementation of the proposed nonparametric regression model using some numerical examples. The data we use include a simulated dataset with a continuous covariate and real data from two oncology studies. As the reversible jump MCMC method is computationally intensive we do not run a full scale Monte Carlo simulation where data sets are generated repeatedly and models are fitted subsequently. the simulated data set is used merely as a proof of concept.

3.3.1 A simulated dataset

Consider data generated from a log-normal nonlinear heteroscedastic model

$$\log T = \mu(Z) + \varepsilon,$$

where the mean function $\mu(Z) = \cos(\pi Z)$ and $\varepsilon|Z \sim N(0, \sigma(Z))$ with $\sigma(Z) = |Z|$. For simplicity the covariate was generated as $Z \sim Unif(0, 1)$ and the censoring time C was generated from an exponential distribution. We set the sample size $n = 100$. We also adjusted the rate parameter for the censoring time in different scenarios to approximately control the censoring rate to take values $\{0, 0.30, 0.50\}$, which represent completely observed, moderately censored, and severely censored data scenarios, respectively. Similar to the previous chapter, except

the Bayesian model based on Bernstein polynomials we also apply the HARE model (Hazard Regression, Kooperberg et al. 1995) to the generated dataset. For the hyperparameters in the Bayesian model we specify $a = 0$, $b = 10$, and $p_0 = 0.5$ for the prior distributions $\pi_\alpha(\cdot; a, b)$ and $\pi_\delta(\cdot; p_0)$. Such specifications are based on the assumption that there is no reliable prior information is available. When there is reliable prior information such as expert opinion is available, one can allow for more meaningful priors for α's and δ's. Note that $\pi_\alpha(\cdot; a, b)$ is the log-normal density with mean a and standard deviation b on the log-scale and $\pi_\delta(\cdot; p_0)$ is Bernoulli with the proportion p_0. Also, we set the prior distribution of m to be a discrete uniform with lower bound $[n^{2/5}]$ and upper bound $[n^{2/3}]$, where $[\cdot]$ denotes rounding up to the nearest integer. As discussed in the previous chapter, these two values correspond to asymptotic rates of convergence. The prior distribution of $\tilde{m}$ is also assumed to be discrete uniform over 5 to 10. This range of the order of Bernstein polynomials for the covariate is rather arbitrary. The standard deviation for the proposal log-normal distribution in the parameter-updating step is set to $\sigma_0 = 0.5$. For the initial values we set $m^{(0)} = \left[\frac{n^{2/5}+n^{2/3}}{2}\right]$, $\tilde{m}^{(0)} = \left[\frac{5+10}{2}\right] = 8$, and $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\alpha}^{(0)} = \boldsymbol{\delta}^{(0)} = \mathbf{1}_{k^{(0)}\times 1}$, where $k^{(0)} = m^{(0)}\tilde{m}^{(0)}$. Finally, we use 55000 iterations with 5000 burn-in's.

The estimates for the survival curves when $Z = 0.5$ obtained by using the proposed Bayesian model and the HARE model are displayed in Figures 3.1, 3.4, and 3.7. Across all censoring scenarios, the estimates provided the Bayesian model based on Bernstein polynomials are very close to the true curves. Although it is only a comparison from one dataset, it appears that the Bayesian model performs slightly better than the HARE model for this particular case. The posterior distributions of m and $\tilde{m}$ are shown in Figure 3.2, 3.5, and 3.8. When the censoring rate is 0, the posterior modes of m and $\tilde{m}$ are generally close to the upper end of the range. When the censoring rate is 30% or 50%, the mode for $\tilde{m}$, the order of Bernstein polynomials for the function of the covariate, jumps to lower values at 7 and 8, respectively.

One challenge when using the reversible jump MCMC sampler is that it is hard to determine the convergence. In literature there is still lack of established procedures to determine the

convergence of reversible jump MCMC. One way to examine the convergence is to look at the trace plots of parameters. Because the dimension of the parameter space is changing along the chain, it is not possible to monitor a certain parameter. Instead, we monitor the trace for the values of the survival function at certain grid points. Specifically, at each iteration of MCMC we compute the survival curve using the current dimension and values of the parameters. The values of the survival function at $t = 0.92, 1.90, 4.00$ are stored. These three grid points represent early, middle, and late stages of a trial or an experiment, respectively. The corresponding trace plots are shown in Figures 3.3, 3.6, and 3.9. It appears the traces are generally well mixed, hence we assume the convergence is approximately achieved for the target distribution.

3.3.2 Gastric cancer data

Recall in the gastric cancer study (Stablein et al., 1981) $n = 90$ patients with locally advanced gastric carcinoma were randomized to two treatment groups (45 patients per group). One group only received chemotherapy while the other group received radiotherapy together with the same chemotherapy. We estimated the survival functions using the proposed Bayesian model. All specifications for the prior distributions and the proposal distributions are the same as the ones used in section 3.3.1. For the initial values we set $m^{(0)} = \left[\frac{n^{2/5}+n^{2/3}}{2}\right]$, $\tilde{m}^{(0)} = \left[\frac{5+10}{2}\right] = 8$, and $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}^{(0)} = \mathbf{1}_{k^{(0)}\times 1}$, where $k^{(0)} = 2m^{(0)}$. Again, we use 55000 iterations with 5000 burn-in's. As shown in Figure 3.10, the estimated survival curves obtained by the proposed model are generally consistent with the ones obtained by the Kaplan-Meier estimator.

The frequency of m in the posterior sample is displayed in Figure 3.11. It appears that the reversible jump sampler tends to go to larger values in the range. For this exaple, we also display the trace plot for m in Figure 3.12. Finally, the trace plots for the values of survival function in Figure 3.13 show that the chains are generally well mixed for the points monitored.

3.3.3 Veteran's Administration lung cancer data

Now we revisit Veteran's administration lung cancer data to demonstrate the use of the proposed Bayesian model when there is a continuous covariate. Recall in this study $n = 137$ male patients were assigned to two treatment groups: standard chemotherapy and test chemotherapy. As discussed in the previous chapter, the important covariates include Karnofsky performance score and cell type. Karnofsky score is a continuous variable that takes value from 0 to 100 while cell type is a categorical variable with fours levels: squamous cell, small cell, adenocarcinoma and large cell.

Similar to the previous example, all specifications for the prior distributions and the proposal distributions are the same as the ones used in section 3.3.1. For the initial values we set $m^{(0)} = \left[\frac{n^{2/5}+n^{2/3}}{2}\right]$, $\tilde{m}^{(0)} = \left[\frac{5+10}{2}\right] = 8$, and $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}^{(0)} = \mathbf{1}_{k^{(0)}\times 1}$, where $k^{(0)} = m^{(0)}(8 + \tilde{m}^{(0)})$. The estimated survival contours are shown for each combination of treatment and cell type in Figure 3.14. The patterns displayed in Figure 3.14 are similar to the ones in the previous chapter. Karnofsky scores have positive association with survival probabilities. It also appears that cell type is a very important factor for the treatment differences. The most significant treatment difference is observed for the patients with small cell lung cancer. For this category, the patients' survival probabilities deteriorate faster in the test chemotherapy treatment group.

Figure 3.15 shows the frequency of m and $\tilde{m}$ in the posterior samples. Most of the selected moves for m and $\tilde{m}$ are around the upper values in the range. We also dsiplay the trace plots for m and $\tilde{m}$ in Figure 3.16. To examine the convergence we select one point at each treatment and cell type combination and create a trace plot of the value the survival function at this point. In total there are 8 trace plots representing 8 categories in Figure 3.17. It appears that the chains are well mixed at least for the points monitored.

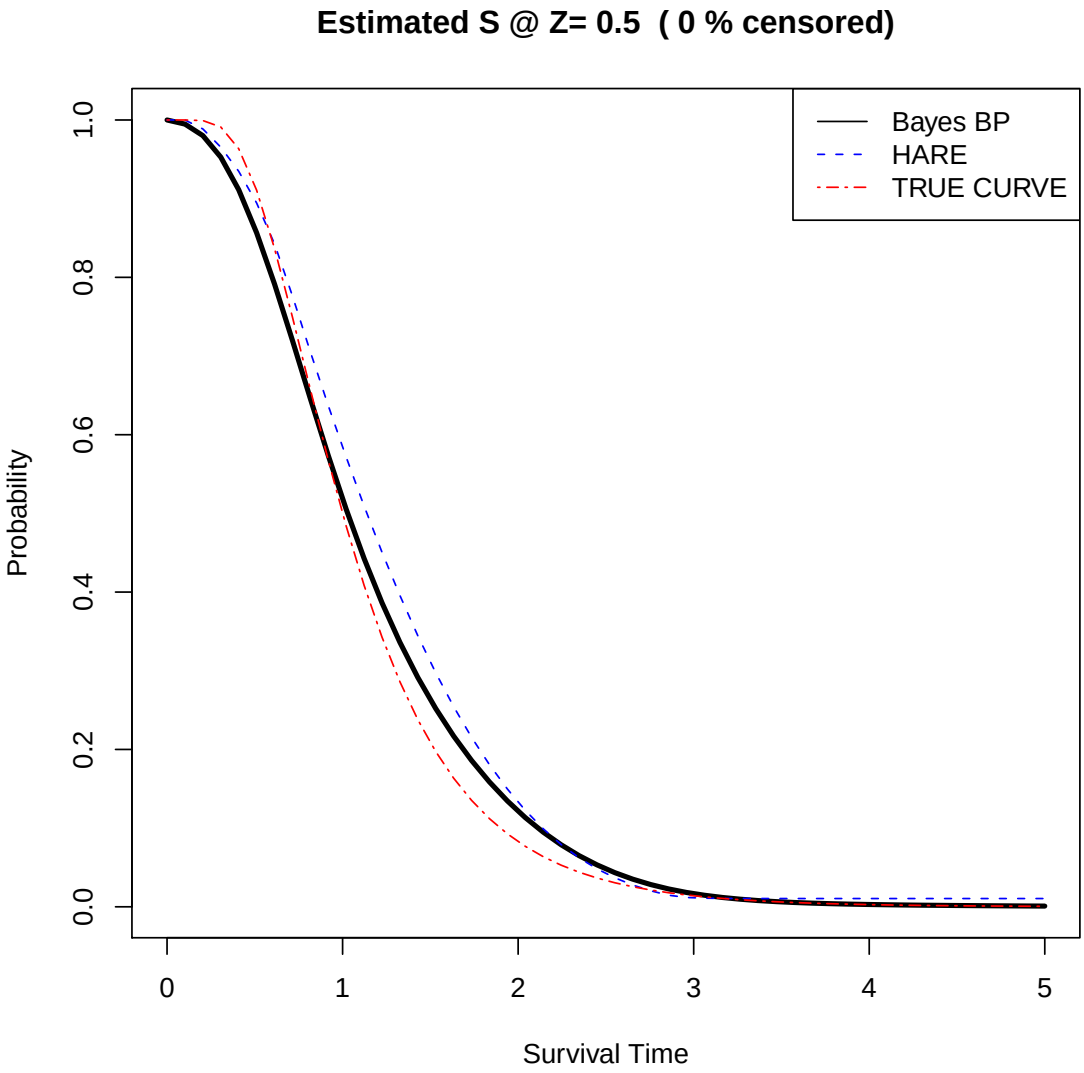
3.4 Conclusions

In this chapter we present a Bayesian implementation of the nonparametric regression model based on Bernstein polynomials for right-censored data. The greatest advantage introduced by using a Bayesian approach is its capability of modeling the uncertainty about the order of Bernstein polynomials using reversible jump Markov Chain Monte Carlo (RJ-MCMC). This has been a great challenge for the proposed method from a frequentist perspective because of the lack of cross-validation procedures for censored data in regression settings. We apply the proposed Bayesian approach to a simulated dataset and data from two cancer studies.

One limitation of the proposed Bayesian model is that the whole process is time consuming. For example, the numerical example with gastric cancer data takes about 65 minutes. Because the slow convergence is expected for the reversible jump sampler, we used long chains in our numerical examples. Besides, even after the sampling step, it takes some time to construct the estimated survival functions from the posterior samples of parameters due to the varying dimensions. So in future work focus should be put on speeding up the sampling and post-processing steps.

In the proposed nonparametric regression model, the methodology is developed only for right-censored data. Due to the flexibility of Bayesian approaches in complex censoring scenarios, the Bayesian nonparametric regression model can be further extended for data subject to various types of interval censoring. Another important extension is the regression models for multivariate time-to-event data. Particularly, the bivariate case is most of interest in practice. For bivariate survival models, the censoring mechanism can be fairly complex. Possible situations include both variables being right-censored, both variables being interval-censored, and the hybrid censoring case with one variable is right-censored and the other is subject to some type of interval censoring such as current status data. These scenarios pose great challenges and yet great opportunities for statistical modeling and inferences. Furthermore, the proposed Bayesian model can be extended to cases with high-dimensional covariates. The current solution for multiple continuous covariates is through an additive structure. Alternatively, a single-index

model can be applied for dimension reduction. However, both of these approaches ignore the possible interactions between covariates. Alternatively, one can incorporate an unsupervised learning technique, such as principle components, to handle high-dimensional covariates.



Note: *BayesBP* (Bayesian regression model based on Bernstein polynomial); *HARE* (Hazard Regression); The red line denotes the true curve

Figure 3.1: Estimated survival curves for the simulated dataset when censoring rate= 0%

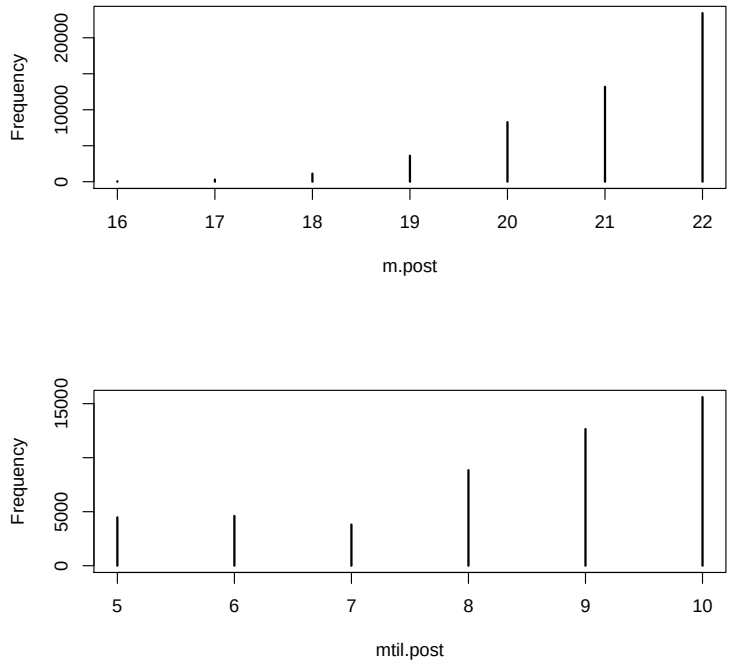


Figure 3.2: Posterior samples for the orders m and $\tilde{m}$ for the simulated dataset when censoring rate= 0%

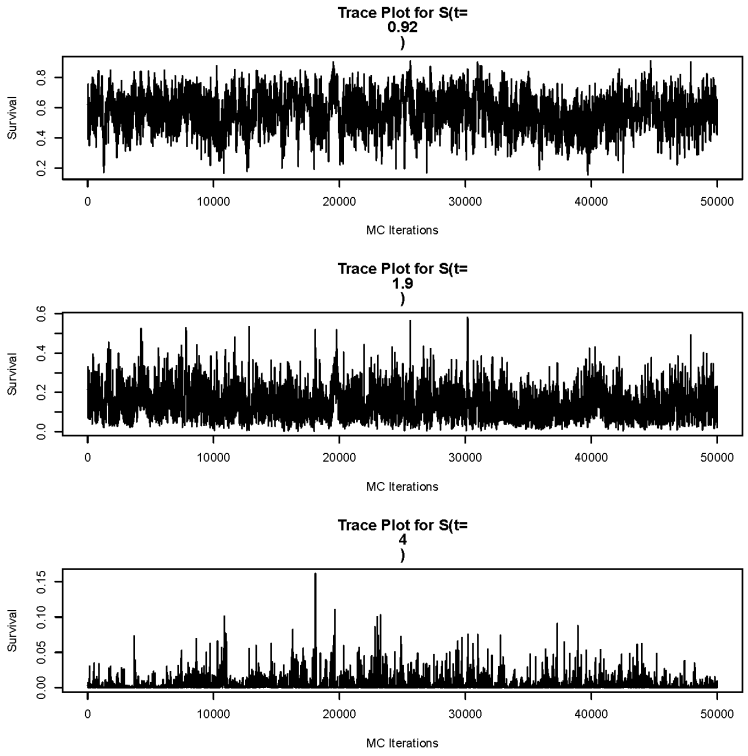
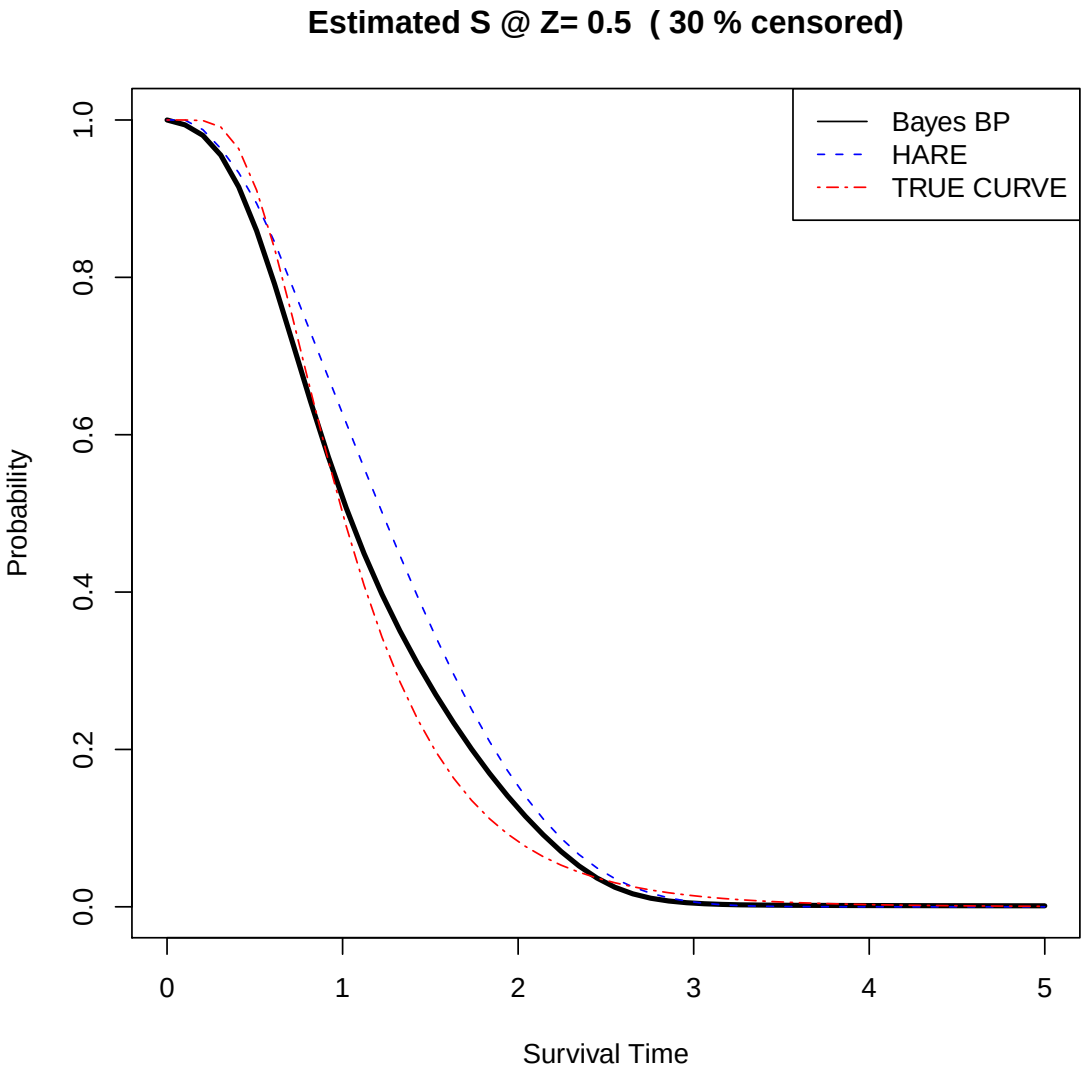


Figure 3.3: Trace plots for the values of survival function at selected time points for the simulated dataset when censoring rate= 0%



Note: *BayesBP* (Bayesian regression model based on Bernstein polynomial); *HARE* (Hazard Regression); The red line denotes the true curve

Figure 3.4: Estimated survival curves for the simulated dataset when censoring rate= 30%

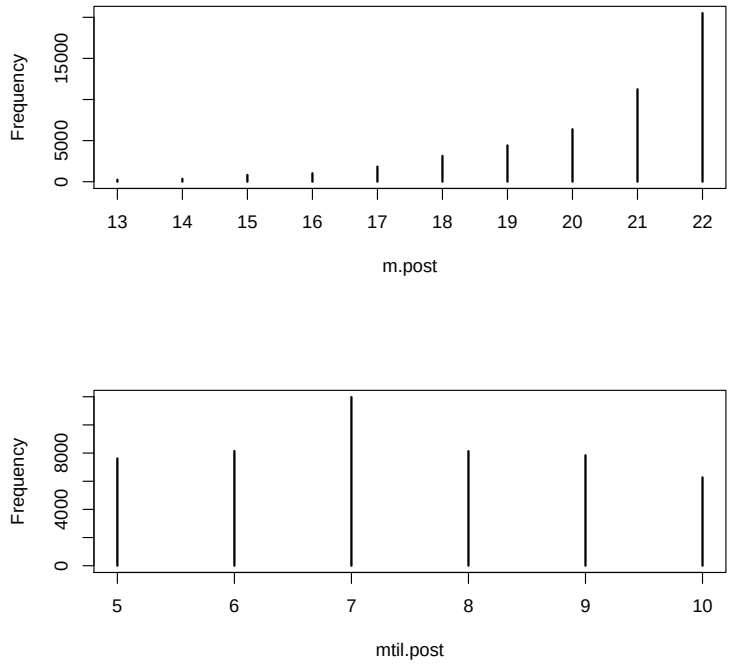


Figure 3.5: Posterior samples for the orders m and $\tilde{m}$ for the simulated dataset when censoring rate= 30%

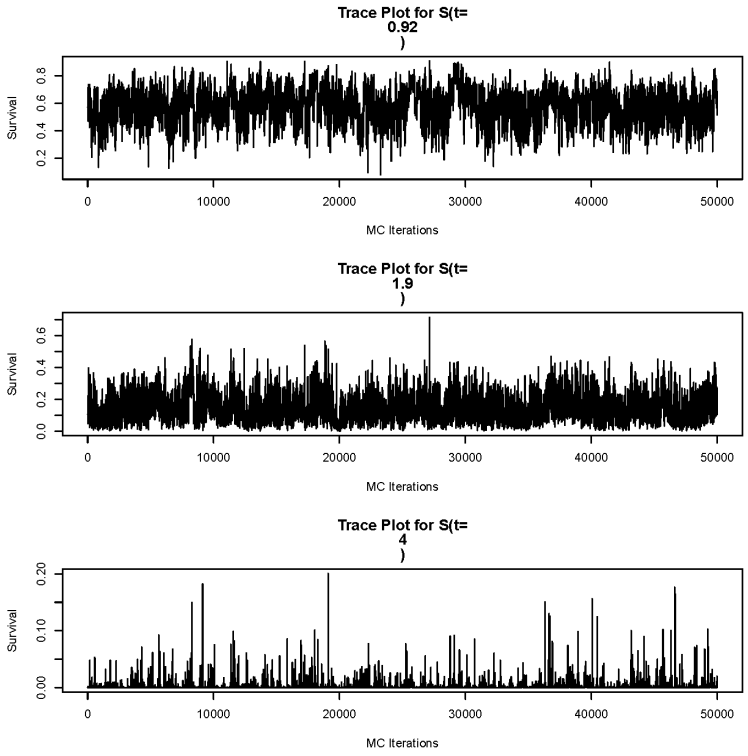
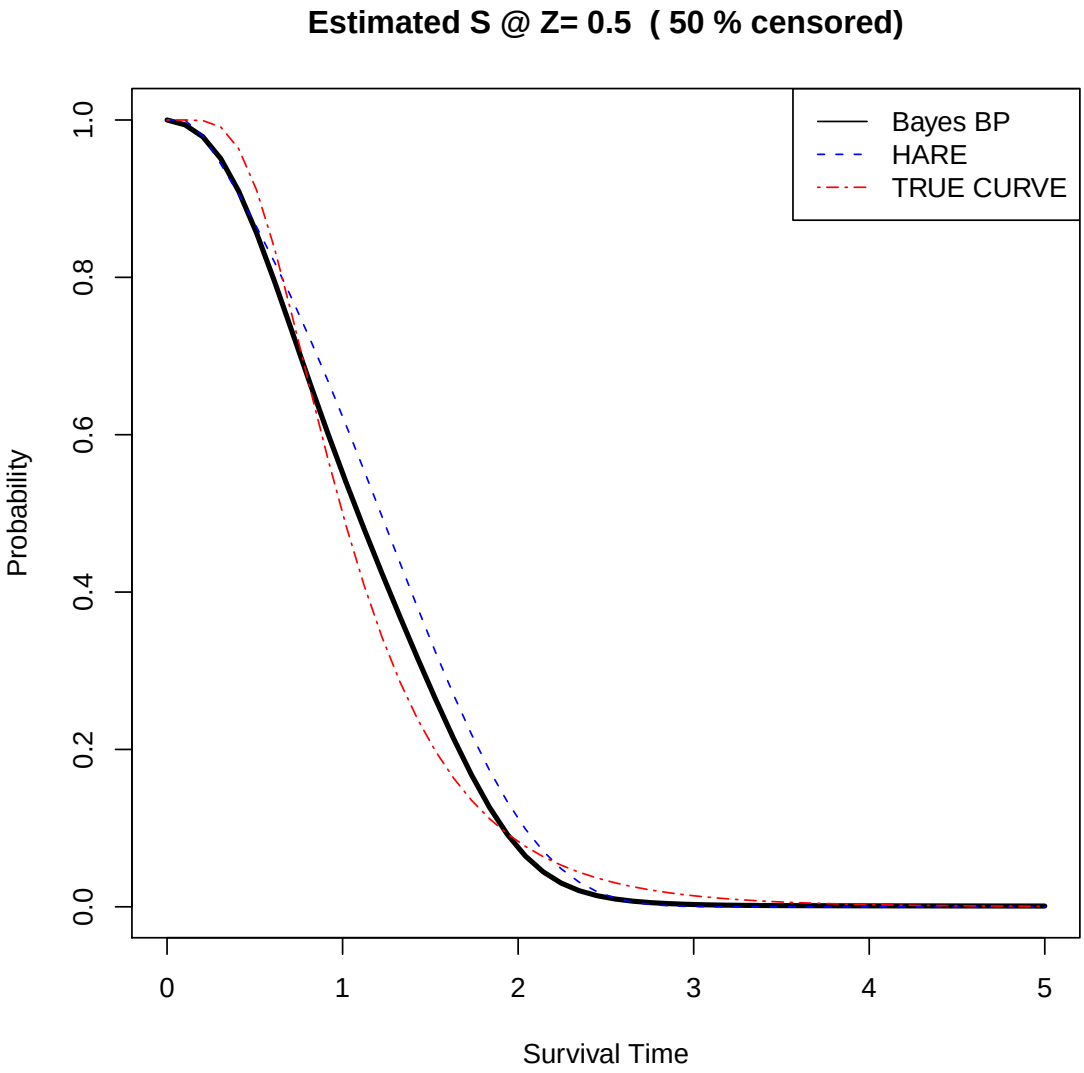


Figure 3.6: Trace plots for the values of survival function at selected time points for the simulated dataset when censoring rate= 30%



Note: *BayesBP* (Bayesian regression model based on Bernstein polynomial); *HARE* (Hazard Regression); The red line denotes the true curve

Figure 3.7: Estimated survival curves for the simulated dataset when censoring rate= 50%

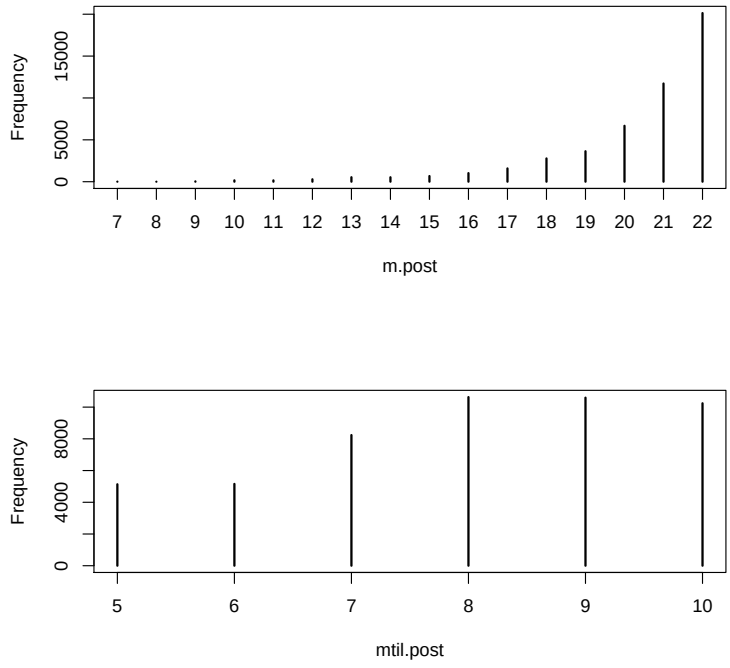


Figure 3.8: Posterior samples for the orders m and $\tilde{m}$ for the simulated dataset when censoring rate= 50%

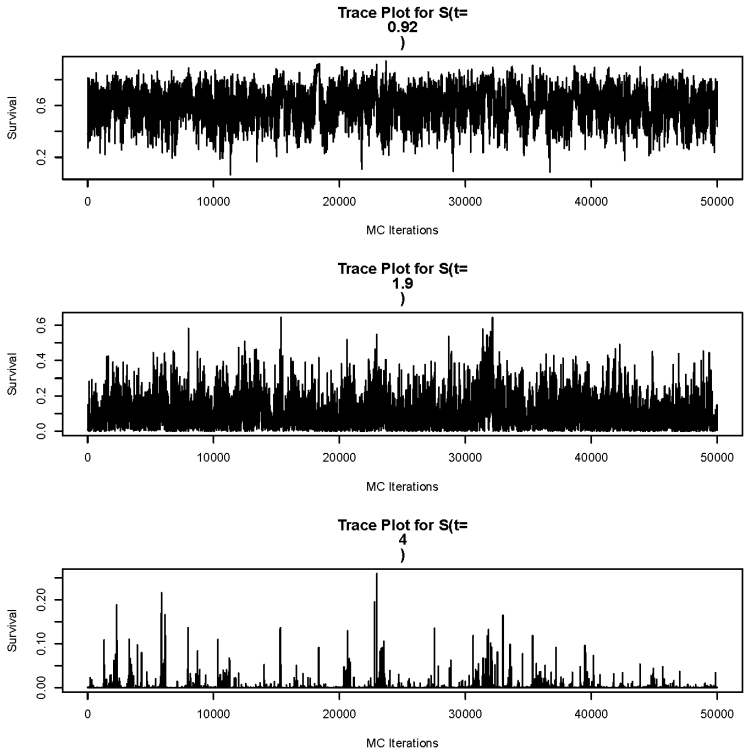


Figure 3.9: Trace plots for the values of survival function at selected time points for the simulated dataset when censoring rate= 50%

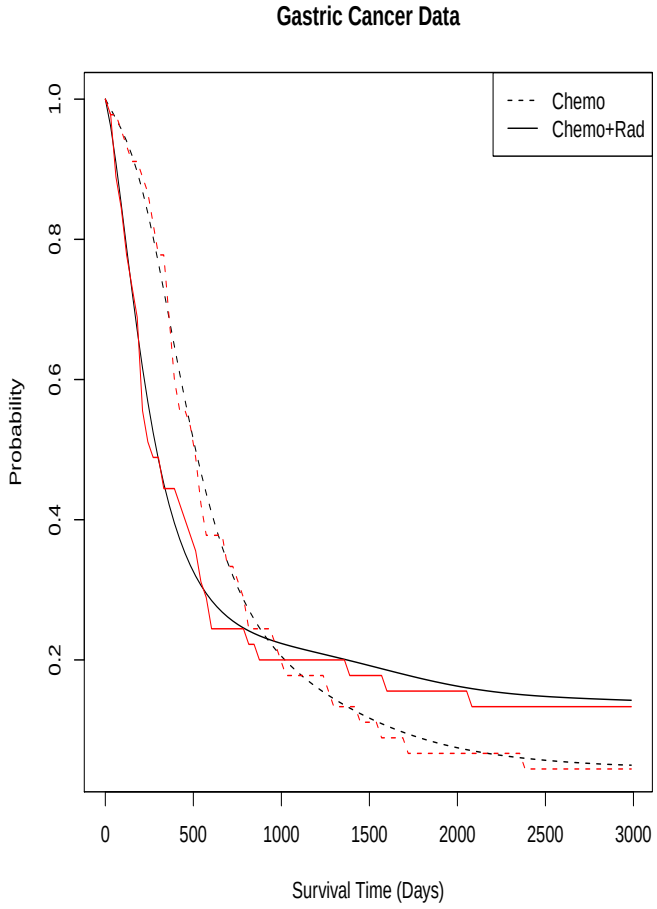


Figure 3.10: Estimated survival curves for the gastric cancer data using the Bayesian model based on Bernstein polynomials and Kaplan-Meier estimator (red)

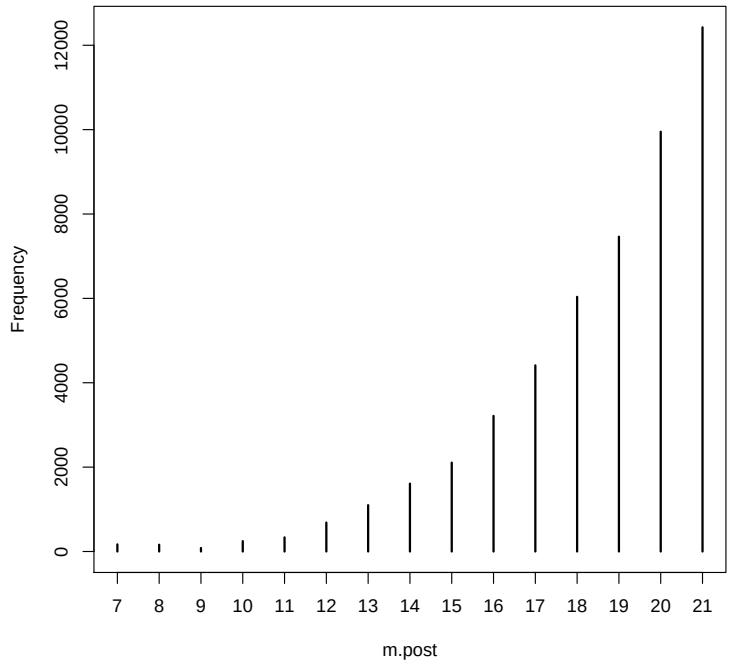


Figure 3.11: Posterior samples for the order m for the gastric cancer data

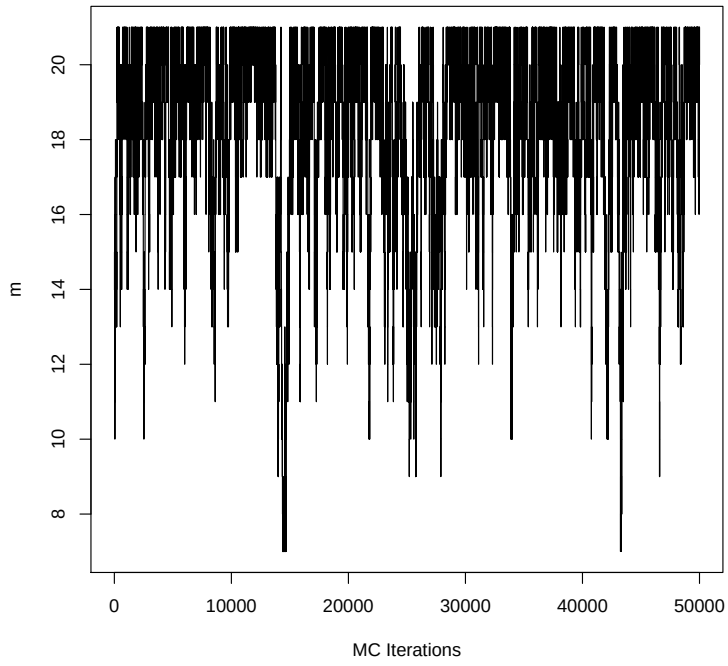


Figure 3.12: Trace plot for the order m for the gastric cancer data

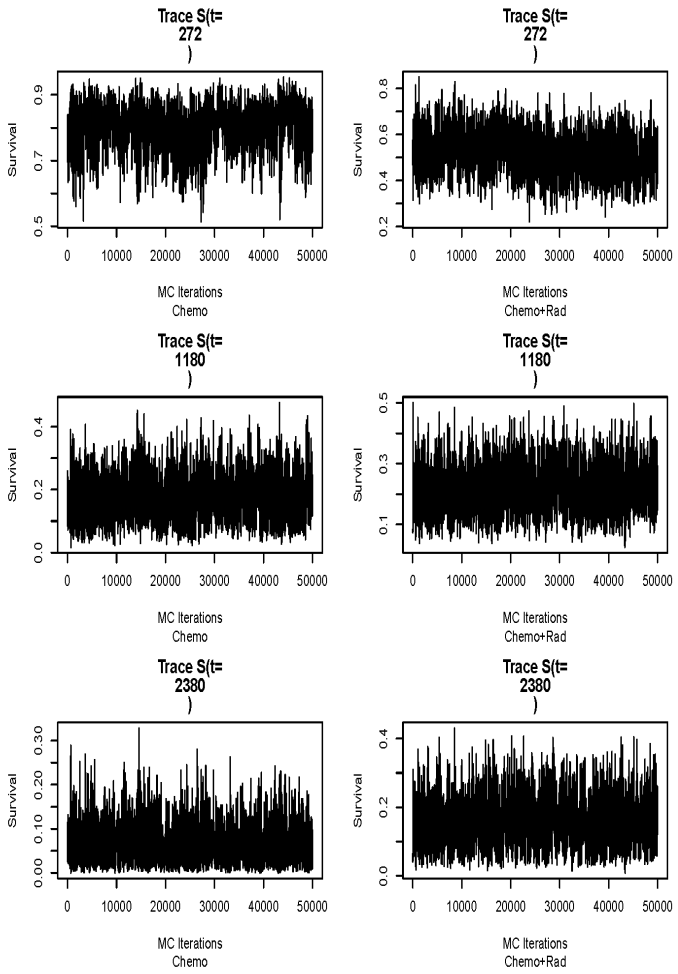
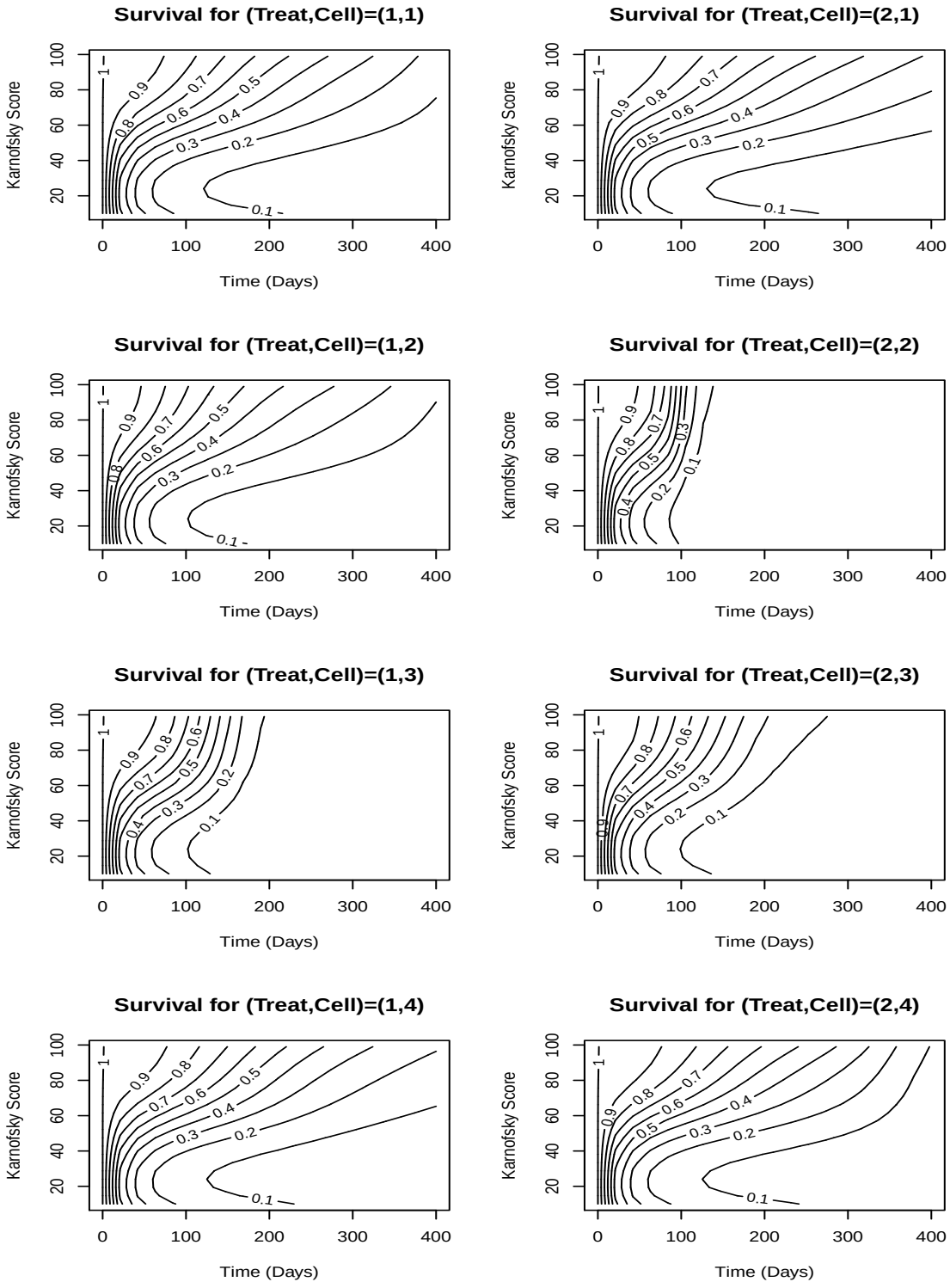


Figure 3.13: Trace plots for the values of survival function at selected time points for the gastric cancer data



Note: 'Treat" denotes treatment group: 1=standard, 2=test; "Cell" denote cell type: 1=squamous, 2=smallcell, 3=adeno, 4=large.

Figure 3.14: Estimated survival contours for the Veteran's Administration lung cancer data

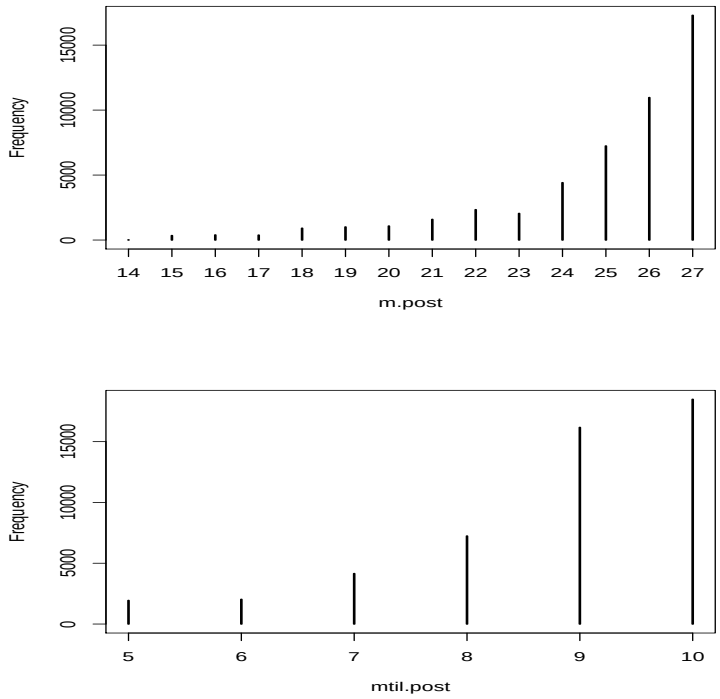


Figure 3.15: Posterior samples for the orders m and $\tilde{m}$ for the Veteran's Administration lung cancer data

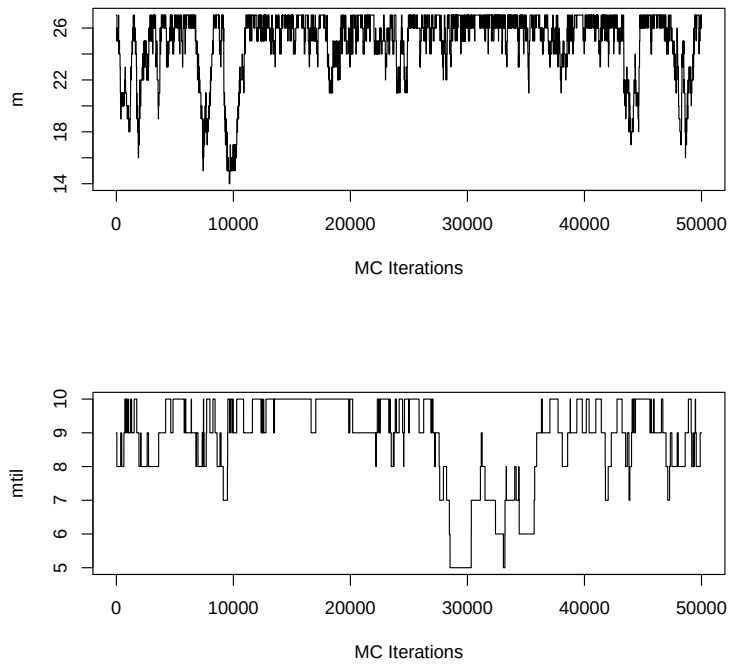


Figure 3.16: Trace plots for m and $\tilde{m}$ for the Veteran's Administration lung cancer data

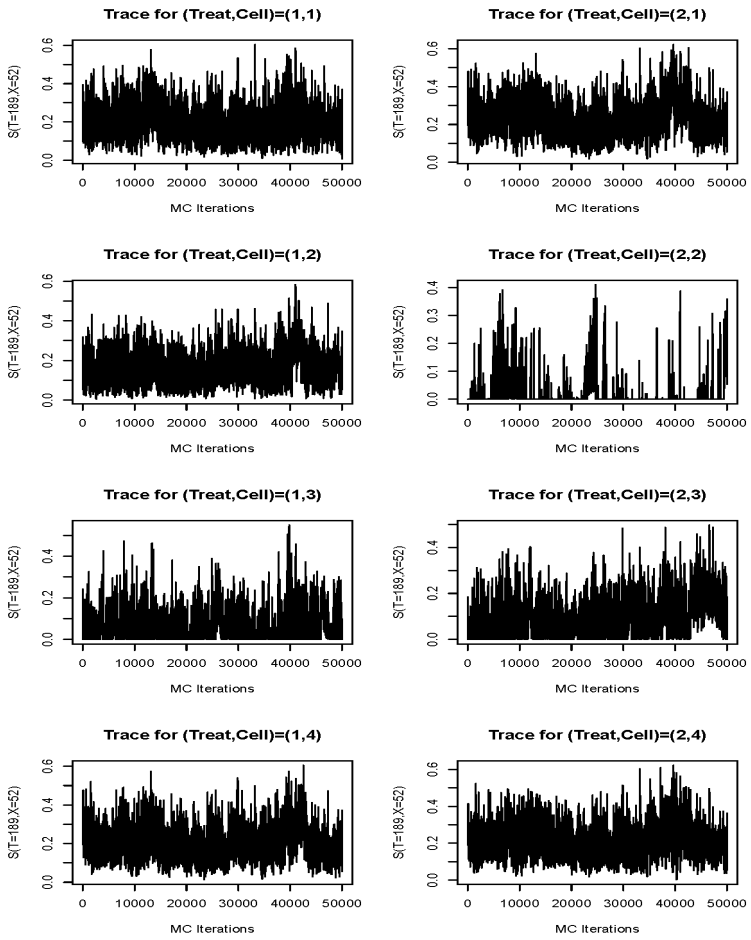


Figure 3.17: Trace plots for the values of survival function at selected time points for the Veteran's Administration lung cancer data

REFERENCES

Babu, G., Canty, A., and Chaubey, Y. (2002) Application of Bernstein Polynomials for smooth estimation of distrbution and density function, *J. Stat. Plan. Infer.*, **105**,377–392.

Bennett, S. (1983), Analysis of survival data by the proportional odds model,*Stat. Med.*,**2**, 273–277.

Blackwelder, W.C. (1982) "Proving the null hypothesis" in clinical trials, *Controlled Clinical Trials*, **3**,345–353.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995) A limited memory algorithm for bound constrained optimization, *SIAM J. Scientific Computing*, **16**, 1190-1208.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Coxs regression models. *Scand. J. Statist.* **30**, 93–111.

Carnicer, J.M. and Peña, J.M., (1993) Shape preserving representations and optimality of the Bernstein basis. *Adv. Comput. Math.*,**1**, 173-196.

Chan, I. S. F. (2003). Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. *Statistical Methods in Medical Research*,**12**, 37–58.

Chang, I.S., Hsiung, C.A., Wu, Y.J. and Yang, C.C. (2005) Bayesian survival analysis using Bernstein polynomials *Scand. J. Statist.*, **32**, 447–466

Chang, I.S., Chien, L.C., Hsiung, C.A., Wen, C.C. and Wu, Y.J. (2007) Shape restricted regression with random Bernstein polynomials *Complex datasets and inverse problems – Institute of mathematical statistics lecture notes – Monograph Series 54, 187-202 Liu, Regina (ed.), Strawderman, William (ed.) and Zhang, Cun-Hui (ed.)* Institute of Mathematical Statistics (Hayward)

Choudhuri, N., Ghosal, S., and Roy, A. (2004) Bayesian estimation of the spectral density of a time series *J. Am. Stat. Assoc.*,**99**, 1050–1059

Chow, S-C. and Shao J. (2006) On non-inferiority margin and statistical tests in active control trial, *Statistics in Medicine*, **25**, 1101–1113.

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

D'Agostino, R.B., Massaro, J.M., and Sullivan, L. (2003) Non-inferiority trials: design concepts and issuesthe encounters of academic consultants in statistics, *Statistics in Medicine*, **22**, 169–186.

Diaconis, P. and Ylvisaker, D. (1985) Quantifying prior opinion, *Bayesian Statistics 2* ,J. Bernardo et al. eds., North-Holland, Amsterdam,133–156.

Farrington, C.P. and Manning, G. (1990) Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk, *Statistics in Medicine*, **9**, 1447–1454.

Gelfand, A. E., and Smith, A. F.M. (1990) Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398-409

Geman, S., and Hwang, C., (1982) Nonparametric Maximum Likelihood Estimation by The Method of SIEVES, *Ann. Statist.*, **10**,401–414.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distribution and Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling.*J. Am. Stat. Assoc.*, **85**, 398–409

Geweke, J. (1996). Variable selection and model comparison in regression. In Bernando, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 5*, 609–620, Oxford University Press, Oxford, UK.

Ghosal, S.(1997). Convergence rates for density estimations with Bernstein polynomials, *The Annals of Statistics*, *29*, 1264–1280.

Ghosh, J. K., Delampady, M. and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer, New York.

Gray, R.J., (1996) Hazard rate regression using ordinary nonparametric regression smoothers. *J. Comput. Graph. Statist.* **5**, 190–207.

Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711–732

Grenander, U., (1981) *Abstract Inference*, Wiley, New York.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109

Hung, H-MJ., Wang, S-J., and O'Neill, R.T. (2005) A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials, *Biometrical Journal*, **47**, 28–36.

Joseph L., du Berger R., and Bélisle P. (1997) Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, **16**, 769–81.

Kakizawa, Y. (2006) Bernstein polynomial estimation of a spectral density *Journal of Time Series Analysis*,**27**, 253–287

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

Kass, R.E. and Raftery, A.E. (1995) Bayes factors, *Journal of the American Statistical Association*, **90**, 773–795.

Klein, J.P. and Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd Ed., New York: Springer.

Kolmogorov, A.N. and Tikhomirov, V.M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk* **14**, 3–86. [In Russian. English translation in *Amer. Math. Soc. Transl. Ser. 2*,**17**,277–364 (1961).]

Kooperberg, C., Stone, C., and Truong, Y.K. (1995) Hazard Regression, *J. Am. Stat. Assoc.*, **90**, 78–94.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya Ser. B*, **60**, 65–81

Laster, L.L. and Johnson, M.F. (2006) Non-inferiority trials: "the at least as good as" criterion for dichotomous data, *Statistics in Medicine*, **22**, 187–200.

Lee, S.J. and Zelen, M., (2000) Clinical trials and sample size considerations: another perspective (with discussion). *Statist. Sci.*,**15**, 95–110.

Li, G., Doss, H., (1995). An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.* **23**, 787–823.

Lindley D.V.(1997) The choice of sample size, *The Statistician*, *46*, 129–38.

Lorentz, G. G. (1956), *Bernstein polynomials*, New York: Chelsea Publishing Co..

Marin, Jean-Michel and Robert, C. P. (2007) *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, New York: Springer.

Martinussen, T. and Scheike, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika* **89**, 283–298.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N. and Teller, A.H. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092

Murphy, S.A. and Sen, P. K. (1991). Time-dependent coefficients in Cox-type regression model. *Stoch. Proces. Applic.***39**, 153–180.

Ng, T.H. (2008) Noninferiority hypotheses and choice of noninferiority margin, *Statistics in Medicine*, **27**, 5392–5406.

Peng, L.M. and Huang Y.J. (2007) Survival analysis with temporal covariate effects. *Biometrika*, **94**, 719–733.

Perperoglou A, Keramopoullos A, van Houwelingen HC. (2007) Approaches in modelling long-term survival: an application to breast cancer. *Stat. Med.*,**26**, 2666–2685.

Petrone, S., (1999) Bayesian Density Estimation Using Bernstein Polynomials. *Can. J. Stat.*,**27**, 105–126.

Prentice, R.L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika*,**60**, 279–288.

Richardson, S. and Green, P. (1997) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination (with discussion). *Journal of the Royal Statistical Society - Series B*, **59**, 731–758

Röhmel, J. (2005) Problems with Existing Procedures to Calculate Exact Unconditional P-Values for Non-Inferiority/Superiority and Confidence Intervals for Two Binomials and How to Resolve Them, *Biometrical Journal*, **47**, 37–47.

Scaglione, F. (1990) Comparison of the clinical and bacteriological efficacy of clarithromycin and erythromycin in the treatment of streptococcal pharyngitis, *Current Medical Research Opinion*, **12**, 25–33.

Schemper, M. (1992) Cox analysis of survival data with non-proportional hazard functions. *Statistician* **41**, 455-465.

Shen, X., and Wong, W.H. (1994) Convergence Rate of Sieve Estimator, *Ann. Stat.*, **23**,580–615.

Sillanpana, M. J. and Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data.*Genetics*, **148**, 1373–1388

Siqueira, A.L., Whitehead, A., and Todd, S. (2008) Active-control trials with binary data: A comparison of methods for testing superiority or non-inferiority using the odds ratio, *Statistics in Medicine*, **27**, 353–370.

Spiegelhalter, D. J., and Freedman, L. S. (1986), A Predictive Approach to Selecting the Size of a Clinical Trial, Based on Subjective Clinical Opinion, *Statistics in Medicine*, **5**, 1–13.

Spierdijk, L. (2008) Nonparametric conditional hazard rate estimation: A local linear approach.*Comput. Stat. Data. An.*,**52** 2419–2434

Stablein, D. M., Carter, W. H. and Novak, J. W. (1981). Analysis of survival data with non-proportional hazard functions. *Control. Clin. Trials*,**2**, 149–159.

Tenbusch, A. (1994) Two-dimensional Bernstein polynomial density estimators *Metrika*, **41**, 233–253.

Therneau T.M. and Grambsch, P.M. (2000) *Modeling Survival Data: Extending the Cox Model*, New York: Springer.

Tian, L., Zucker, D. and Wei, L. J. (2005). On the Cox model with time-varying regression coefficients. *J. Am. Statist. Assoc.* **100**, 172–183.

Tu, D. (1998) On the use of the ratio or the odds ratio of cure rates in therapeutic equivalence clinical trials with binary endpoints, *Journal of Biopharmaceutical Statistics*, **8**, 263–282.

van de Geer, S. (2000) *Empirical Processes in M-Estimation*, Cambridge Univ. Press.

van Keilegom, I., Veraverbeke, N., (2001). Hazard rate estimation in nonparametric regression with censored data. *Ann. Inst. Statist. Math.* **53**,730–745.

Vitale, Richard A. (1975) A Bernstein polynomial approach to density function estimation. *Stochastic Processes and Related Topics*,**2**, 87–100.

Waagepetersen, R. and Sorensenz, D. (1998). A Tutorial on Reversible Jump MCMC with a View toward Applications in QTL-mapping.*International Statistical Review*, **69**, 49–61

Wellek, S. (2003) *Testing Statistical Hypotheses of Equivalence*. Chapman & Hall/CRC: London.

Wellek, S. (2005) Statistical methods for the analysis of two-arm non-inferiority trials with binary outcomes, *Biometrical Journal*, **47**, 48–61.

Williamson, P.P. (2007) Bayesian equivalence tesing for binomail random variables, *Journal of Statistical Computation and Simulation*, **77**, 739–755.

Weiss, R. (1997) Bayesian Sample Size Calculations for Hypothesis Testing, *The Statistician*, **46**, 185–191.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent effects: a penalized partial likelihood approach. *Ann. Statist.* **18**, 329–353.

APPENDICES

Appendix A

Proof for the upper bound of the avarge type I error

Theorem 1. *There exists* $\zeta_0 \in (0,1)$ *such that for any weight* $\zeta < \zeta_0$ *in* TWE_1 *the Bayesian type I error satisfies*

$$BE_1(B_0) \leq TWE_1(B_0) \leq \zeta,$$

where $B_0 = \frac{1-\zeta}{\zeta}$, $BE_1(B_0) = P[BF(X) > B_0 | \theta \in \Theta_0]$, *and* $TWE_1(B_0)$ *is as defined in (1.8).*

Proof. We establish this result under the general case when the sampling density $f(x|\theta)$ and prior density $\pi(\theta)$ are arbitrary and not necessarily based on binary response data or conjugate priors.

Let us denote the Bayesian type II error as $BE_2(B_0) = P[BF(X) \leq B_0 | \theta \in \Theta_1]$, then the total error can be written as

$$TWE_1(B_0) = (1-\zeta)BE_1(B_0) + \zeta BE_2(B_0).$$

As $TWE_1(B_0)$ is a (linear) convex function of ζ, we have $TWE_1(B_0) > \min[BE_1(B_0), BE_2(B_0)]$. Therefore, the Bayesian type I error $BE_1(B_0)$ is bounded by $TWE_1(B_0)$ if $BE_1(B_0) - BE_2(B_0) \leq 0$.

Since

$$\begin{aligned} BE_1(B_0) - BE_2(B_0) &= P[BF(X) > B_0|\theta \in \Theta_0] - P[BF(X) \leq B_0|\theta \in \Theta_1] \\ &= 1 - \int I[BF(X) \leq B_0][BF(X)+1]m_0(X)dX \\ &= 1 - \int I[BF(X) \leq \frac{1-\zeta}{\zeta}][BF(X)+1]m_0(X)dX \\ &= \Delta_{BE}(\zeta) \end{aligned}$$

where $m_0(X) = \int_{\Theta_0} f(X|\theta)\pi_0(\theta)d\theta$ and $\pi_0(\theta) = \frac{\pi(\theta)I[\theta\in\Theta_0]}{P[\theta\in\Theta_0]}$. Clearly, $\Delta_{BE}(\zeta)$ is a monotonic increasing function of ζ with $\lim_{\zeta\to 0}\Delta_{BE}(\zeta) < 0$ and $\lim_{\zeta\to 1}\Delta_{BE}(\zeta) > 0$. So there exists a $\zeta_0 = \Delta_{BE}^{-1}(0) = \sup\{\zeta \in (0,1) : \Delta_{BE}(\zeta) \leq 0\}$. It follows that $\Delta_{BE}(\zeta) = BE_1(B_0) - BE_2(B_0) \leq 0$ if $\zeta \leq \zeta_0$. Therefore, the Bayesian type I error $BE_1(B_0)$ is bounded by $TWE_1(B_0)$ if $\zeta \leq \zeta_0$.

The second part of the inequality is rather straightforward because $TWE_1(B_0)$ is the minimum value over all cut-off values and when the cut-off goes to infinity the total error becomes ζ. As a result, $TWE_1(B_0) \leq \zeta$. This completes the proof. □

Remark 1: Generally, the condition $\zeta \leq \zeta_0$ holds for most practical scenarios because the type I error is usually given much larger weight than the type II error. To obtain the value of ζ_0, one can essentially solve the equation $\Delta(\zeta) = 0$. Even though there is no analytical solution for most cases, it should be relatively trivial to compute ζ_0 using software packages such as "uniroot" in R. In the noninferiority testing problem for binary data in particular, the right-hand side of the equation becomes $\Delta(\zeta) = 1 - \sum_{x_1=0}^{n_1}\sum_{x_2=0}^{n_2} I[BF(x_1,x_2) \leq \frac{1-\zeta}{\zeta}][BF(x_1,x_2)+1]P(x_1,x_2)\frac{Pr[(\theta_1,\theta_2)\in\Theta_0|X_1=x_1,X_2=x_2]}{Pr[(\theta_1,\theta_2)\in\Theta_0]}$, where $BF(x_1,x_2)$, $P(x_1,x_2)$, $Pr[(\theta_1,\theta_2) \in \Theta_0|X_1 = x_1, X_2 = x_2]$, and $Pr[(\theta_1,\theta_2) \in \Theta_0]$ are defined in (1.6), (1.7), (1.5), and (1.3), respectively. For the numerical examples in this paper, the value of ζ_0 is approximately 0.45 for the simulation studies and it is almost exactly 0.5 for the real data example.

Appendix B

Proofs of consistency and rate of convergence

The proof of consistency is built within the framework of Theorem 1 of Geman and Hwang (1982). Following Geman and Hwang (1982), the following notations and definitions will be used: 1) For $h \in \Theta_m$ and $\forall \varepsilon > 0$, an open ball is defined as $B_m(h, \varepsilon) = \{g : g \in \Theta_m \text{ and } d(h, g) < \varepsilon\}$; 2) $E_h g(X) = \int g(x) p_h(x) d\mu(x)$ and $Q(h, g) = E_h \log p_g(X)$; 3) For a function $g : \Theta \to \mathbb{R}, g(B) = \sup_{b \in B} g(b)$ for any $B \subseteq \Theta$; 4) The maximum entropy set is defined as $A_m = \{h : h \in \Theta_m \text{ and } Q(h_0, h) = Q(h_0, \Theta_m)\}$. 5) For $C_m \subset A$, $C_m \to h$ means $\sup_{g \in C_m} d(h, g) \to 0$.

Lemma 1. *(Geman and Hwang, 1982) Assume the sieve $\{\Theta_m\}$ satisfies following conditions:*

B1. $f(x, B_m(h, \varepsilon))$ is measurable in x for $\forall m, \forall h \in \Theta_m, \forall \varepsilon > 0$ and $f(x, h)$ is upper-semicontinuous in h on Θ_m;

B2. $\exists \varepsilon > 0$ such that $E_{h_0} \log f(X, B_m(h, \varepsilon)) < \infty$ for $\forall m, \forall h \in \Theta_m$;

B3. Θ_m is compact for $\forall m$;

B4. $A_m \to h_0$ as $m \to \infty$,

Then the sieve maximum likelihood estimator $\hat{h}_{m_n}(\cdot) \to_{a.s.} h_0(\cdot)$ as $m, n \to \infty$.

Proof of Theorem 3.1 (Consistency): Since the function $f(x,\delta,\theta) = \theta(x)^{\delta} e^{-\int_0^x \theta(u)du}$ is continuous with respect to $\theta(\cdot)$, it immediately follows that B1 is satisfied. Also, when $\theta \in \Theta_m$, the function $\theta(x,\boldsymbol{\gamma}_m)$ can be regarded as a continuous mapping from $[0,L_\gamma]^m$ to Θ_m. Given $[0,L_\gamma]^m$ is a compact space, it follows that Θ_m is also compact, so B3 holds.

For any $h \in \Theta_m$, it is easy to verify the discrepancy $D(h_0,h) = E_0\left\{\log\frac{f(X,\Delta,h_0)}{f(X,\Delta,h)}\right\} \geq 0$ holds for right-censored data. So we have

$$\begin{aligned} E_0 \log f(X,\Delta,h) &\leq E_0 \log f(X,\Delta,h_0) \\ &= E_0 \log\left\{h_0(X)^{\Delta} S_0(X)\right\} \\ &\leq E_0 \Delta \log h_0(X). \end{aligned}$$

Since $h_0(X)$ is bounded and bounded away from 0 by the condition (II), we have $E_0 \log f(X,\Delta,h) < \infty$ for $\forall h \in \Theta_m$. This shows that B2 is satisfied.

Let $\tilde{h}_m(t) = \sum_{k=1}^m h_0(\frac{k}{m}\tau)\binom{m-1}{k-1}(t/\tau)^{k-1}(1-t/\tau)^{m-k}$, clearly $\tilde{h}_m \in \Theta_m$. And from Bernstein-Weierstrass theorem (Lorentz, 1956), $||h_0(\cdot) - \tilde{h}_m(\cdot)||_\infty = \sup_{0\leq t\leq\tau}|h_0(t) - \tilde{h}_m(t)| \to 0$ as $m \to \infty$. The discrepancy between $f_(x,\delta,h_0)$ and $f_(x,\delta,h_m)$

$$\begin{aligned} D(h_0,\tilde{h}_m) &= E_0 \log\frac{f(X,\Delta,h_0)}{f(X,\Delta,\tilde{h}_m)} \\ &= E_0 \Delta \log\frac{h_0(X)}{\tilde{h}_m(X)} + E_0\int_0^X [h_0(u) - \tilde{h}_m(u)]du \\ &\leq E_0 \Delta \log\left\{1 + \frac{h_0(X) - \tilde{h}_m(X)}{\tilde{h}_m(X)}\right\} + E_0\{\tau||h_0(\cdot) - \tilde{h}_m(\cdot)||_\infty\} \\ &\leq E_0 \Delta \frac{h_0(X) - \tilde{h}_m(X)}{h_0(X)} + \tau||h_0(\cdot) - \tilde{h}_m(\cdot)||_\infty \\ &\leq E_0 \frac{|h_0(X) - \tilde{h}_m(X)|}{h_0(X)} + \tau||h_0(\cdot) - \tilde{h}_m(\cdot)||_\infty \\ &\leq E_0 \frac{||h_0(\cdot) - \tilde{h}_m(\cdot)||_\infty}{h_0(X)} + \tau||h_0(\cdot) - \tilde{h}_m(\cdot)||_\infty \\ &= (E_0 \frac{1}{h_0(X)} + \tau)||h_0(\cdot) - \tilde{h}_m(\cdot)||_\infty. \end{aligned} \tag{B.1}$$

Again, due to the fact that h_0 is bounded away from 0 (see the condition (II)) and $\sup_{0\le u\le\tau}|h_0(u)-\tilde{h}_m(u)|\to 0$ as $m,n\to 0$, there exists $\tilde{h}_m\in\Theta_m$ with $D(h_0,\tilde{h}_m)\to 0$, this implies that B4 is satisfied. Finally, by *Remark 1* in Section 2, it follows that the set consists of sieve maximum likelihood estimators is non-empty. By Lemma 1, $\hat{h}_m(\cdot)\to_{a.s} h_0(\cdot)$ as $m,n\to\infty$. This completes the proof of consistency.

The proof of the converegence rate is built on a result given by Shen and Wong (1994).

Lemma 2. *(Shen and Wong, 1994) For the sieve maximum likelihood estimator*

$$\hat{h}_m=\arg\max_{h\in\Theta_m}\sum_{i=1}^{n}l(X_i,\Delta_i,h),$$

if following two conditions hold:

C1.$\inf_{\{d(h_0,h_m)\ge\varepsilon\}}E_0\{l(X,\Delta,h_0)-l(X,\Delta,h_m)\}\ge c_2\varepsilon^{2\alpha}$*;*

C2.$\sup_{\{d(h_0,h_m)\le\varepsilon\}}VAR_0\{l(X,\Delta,h_0)-l(X,\Delta,h_m)\}\le c_3\varepsilon^{2\beta}$,

additionally,

C3. for function class $\mathcal{F}_m=\{l(h,\cdot)-l(\tilde{h}_m,\cdot):h\in\Theta_m\}$ *the uniform metric entropy satisfies* $H(\varepsilon,\mathcal{F}_m)\le c_4n^{2r_0}\varepsilon^{-r}$ *for the case of* $0^+\le r<2$*, then*

$$d(\hat{h}_m,h_0)=O_p(\max(n^{-\tau_0},d(\pi_nh_0,h_0),D^{1/2\alpha}(\pi_nh_0,h_0))),$$

where $D(\pi_nh_0,h_0)=E_0(l(X,\Delta,h_0)-l(X,\Delta,\pi_nh_0))$*,* π_nh_0 *is an approximation of* h_0 *in the sieve* Θ_m*, and*

$$\tau_0=\begin{cases}\frac{1-2r_0}{2\alpha}-\frac{\log\log n}{2\alpha\log n} & \text{if } r=0^+,\beta\ge\alpha\\ \frac{1-2r_0}{4\alpha-2\beta} & \text{if } r=0^+,\beta<\alpha\\ \frac{1-2r_0}{4\alpha-\min(\alpha,\beta)(2-r)} & \text{if } 0<r<2\\ \frac{1-2r_0}{4\alpha}-\frac{\log\log n}{2\alpha\log n} & \text{if } r=2\\ \frac{1-2r_0}{2\alpha r} & \text{if } r>2.\end{cases}$$

Proof of Theorem 3.2 (Rate of Converdence): For the Bernstein polynomial

$$\tilde{h}_m(t) = \sum_{k=1}^{m} h_0(\frac{k}{m}\tau)\binom{m-1}{k-1}(t/\tau)^{k-1}(1-t/\tau)^{m-k},$$

it follows from the Theorem 1.6.2 of Lorentz (1956) and the condition (II) that the approximation error $|\tilde{h}_m(t) - h_0(t)| \leq \frac{3}{4}(m-1)^{-\frac{1+\alpha_0}{2}} \approx \frac{3}{4}m^{-\frac{1+\alpha_0}{2}}$ when m is large enough. Assume $m = o(n^\kappa)$, we have

$$\sup_{0\leq t\leq\tau} |\tilde{h}_m(t) - h_0(t)| \leq \frac{3}{4}n^{-\frac{(1+\alpha_0)\kappa}{2}}.$$

First, since the discrepancy

$$\begin{aligned}
D(h_0, h_m) &= E_0\left\{\log\frac{f(X,\Delta,h_0)}{f(X,\Delta,h_m)}\right\} \\
&= 2E_0\left\{-\log\sqrt{\frac{f(X,\Delta,h_m)}{f(X,\Delta,h_0)}}\right\} \\
&\geq 2E_0\left\{1-\sqrt{\frac{f(X,\Delta,h_m)}{f(X,\Delta,h_0)}}\right\} \\
&= 2-2\int\sqrt{\frac{f(x,\delta,h_m)}{f(x,\delta,h_0)}}f(h_0,x,\delta)g(h_c,x,\delta)d\mu(x,\delta) \\
&= 2\left\{1-\int\sqrt{f(x,\delta,h_m)g(x,\delta,h_c)}\sqrt{f(h_0,x,\delta)g(h_c,x,\delta)}d\mu(x,\delta)\right\} \\
&= 2d(h_0,h_m)^2,
\end{aligned}$$

the condition C1 is satisfied with $\alpha = 1$.

The condition C2 could be verified to hold with $\beta = 1$ following the similar steps as in the

example 2 of Shen and Wong (1994)

$$
\begin{aligned}
& VAR_0\{l(X,\Delta,h_0) - l(X,\Delta,h_m)\} \\
\leq\ & E_0\{l(X,\Delta,h_m) - l(X,\Delta,h_0)\}^2 \\
\leq\ & E_0\left\{\log\frac{f(X,\Delta,h_m)}{f(X,\Delta,h_0)}\right\}^2 \\
\leq\ & 4E_0\left\{\log\sqrt{\frac{f(X,\Delta,h_m)}{f(X,\Delta,h_0)}}\right\}^2 \\
\leq\ & 4E_0\left\{\log(1+(\sqrt{\frac{f(X,\Delta,h_m)}{f(X,\Delta,h_0)}}-1))\right\}^2 \\
\leq\ & 4BE_0\left\{\sqrt{\frac{f(X,\Delta,h_m)}{f(X,\Delta,h_0)}}-1\right\}^2 \\
=\ & 4BE_0\left\{\frac{f(X,\Delta,h_m)}{f(X,\Delta,h_0)}+1-2\sqrt{\frac{f(X,\Delta,h_m)}{f(X,\Delta,h_0)}}\right\} \\
=\ & 4B\{\int\frac{f(x,\delta,h_m)}{f(x,\delta,h_0)}f(x,\delta,h_0)g(x,\delta,h_c)d\mu(x,\delta)+1 \\
& -2\int\sqrt{\frac{f(x,\delta,h_m)}{f(x,\delta,h_0)}}f(x,\delta,h_0)g(x,\delta,h_c)d\mu(x,\delta)\} \\
=\ & 4B\{2-2\int\sqrt{f(x,\delta,h_m)g(x,\delta,h_c)}\sqrt{f(x,\delta,h_0)g(x,\delta,h_c)}d\mu(x,\delta)\} \\
=\ & 8Bd(h_0,h_m)^2,
\end{aligned}
$$

for some $B \geq 1$ if $\frac{f(X,\Delta,h_m)}{f(X,\Delta,h_0)} \geq \eta > -1$, which is satisfied because $f(X,\Delta,h_0)$ is bounded.

Let $\mathcal{F}_m = \{l(h,\cdot) - l(\tilde{h}_m,\cdot) : h \in \Theta_m\}$, where $\tilde{h}_m(t) = \sum_{k=1}^m h_0(\frac{k}{m}\tau)\binom{m-1}{k-1}(t/\tau)^{k-1}(1-t/\tau)^{m-k}$. Further, let's denote $N(\varepsilon,\mathcal{F}_m)$ as the minimum number of ε-balls in uniform metric required to cover $\mathcal{F}_m$, then $H(\varepsilon,\mathcal{F}_m) = \log N(\varepsilon,\mathcal{F}_m)$ is the L_∞-metric entropy of the space $\mathcal{F}_m$. From Kolmogrov and Tihomirov (1959) and the condition (III), we have

$$H(\varepsilon,\mathcal{F}_m) \leq H(\varepsilon,\Theta_m) \leq c_4\varepsilon^{-1/(1+\alpha_0)}.$$

So the condition C3 is satisfied with $r_0 = 0$ and $r = \frac{1}{1+\alpha_0}$. Consequently,

$$\tau_0 = \frac{1 - 2r_0}{4\alpha - \min(\alpha, \beta)(2 - r)} = \frac{1}{4 - (2 - \frac{1}{1+\alpha_0})} = \frac{1 + \alpha_0}{3 + 2\alpha_0}.$$

Therefore, the rate of convergence

$$\begin{aligned} d(\hat{h}_m, h_0) &= O_p(\max(n^{-\tau_0}, d(\tilde{h}_m, h_0), K^{1/2\alpha}(f(\cdot, \tilde{h}_m), f(\cdot, h_0)))) \\ &= O_p(\max(n^{-\frac{1+\alpha_0}{3+2\alpha_0}}, n^{-\frac{(1+\alpha_0)\kappa}{4}})) \end{aligned}$$

If we let $\kappa = \frac{2}{3+2\alpha_0}$, we have the rate of convergence $d(\hat{h}_m, h_0) = O_p(n^{-\frac{1+\alpha_0}{3+2\alpha_0}})$. Note that α_0 is the exponent of the Holder's continuity of $h_0^{(1)}$. If $\alpha_0 = 1$, h_0 has finite second derivative, then the rate of convergence $d(\hat{h}_m, h_0) = O_p(n^{-2/5})$ with $\kappa = 2/5$. On the other hand, h_0 only has finite first derivative when $\alpha_0 = 0$, this in turn gives $d(\hat{h}_m, h_0) = O_p(n^{-1/3})$ with $\kappa = 2/3$.