2012

Semiparametric Survival Analysis Using Models with Log-Linear Median

Jianchang Lin

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

SEMIPARAMETRIC SURVIVAL ANALYSIS USING MODELS WITH LOG-LINEAR MEDIAN

By

JIANCHANG LIN

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 2012

Jianchang Lin defended this dissertation on March 26, 2012.

The members of the supervisory committee were:

Debajyoti Sinha
Professor Directing Thesis

Yi Zhou
University Representative

Stuart Lipsitz
Committee Member

Dan McGee
Committee Member

Xu-Feng Niu
Committee Member

Yiyuan She
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with the university requirements.

To my wife Fang, my grandfather Jinbao Lin and my parents

ACKNOWLEDGMENTS

First of all, I would like to take this opportunity to express special thanks to my advisor, Dr. Debajyoti Sinha for inspiring me to study Bayesian Biostatistics, and for all of his invaluable guidance, advice and support during my graduate career in the FSU. I could not have asked for a better advisor than Dr. Sinha. I greatly appreciates the time and faith that he invested in me as a student.

I am also thankful to Dr. She, Dr. Polpo, Dr. Lipsitz for their collaboration, suggestions and help on my research work. My special thanks also go to committee members Dr. Niu and Dr. McGee and their wonderful lectures. Thanks for the support and encouragement on my dissertation.

I am also very grateful to all the faculty members and all the staffs in the Department of Statistics for their generous help during my time at Florida State University.

Thank you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

First, we present two novel semiparametric survival models with log-linear median regression functions for right censored survival data. These models are useful alternatives to the popular Cox (1972) model and linear transformation models (Cheng et al., 1995). Compared to existing semiparametric models, our models have many important practical advantages, including interpretation of the regression parameters via the median and the ability to address heteroscedasticity. We demonstrate that our modeling techniques facilitate the ease of prior elicitation and computation for both parametric and semiparametric Bayesian analysis of survival data. We illustrate the advantages of our modeling, as well as model diagnostics, via reanalysis of a small-cell lung cancer study. Results of our simulation study provide further guidance regarding appropriate modelling in practice.

Our second goal is to develop the methods of analysis and associated theoretical properties for interval censored and current status survival data. These new regression models use log-linear regression function for the median. We present frequentist and Bayesian procedures for estimation of the regression parameters. Our model is a useful and practical alternative to the popular semiparametric models which focus on modeling the hazard function. We illustrate the advantages and properties of our proposed methods via reanalyzing a breast cancer study.

Our other aim is to develop a model which is able to account for the heteroscedasticity of response, together with robust parameter estimation and outlier detection using sparsity penalization. Some preliminary simulation studies have been conducted to compare the performance of proposed model and existing median lasso regression model. Considering the estimation bias, mean squared error and other identification benchmark measures, our proposed model performs better than the competing frequentist estimator.

CHAPTER 1

SEMIPARAMETRIC BAYESIAN SURVIVAL ANALYSIS USING MODELS WITH LOG-LINEAR MEDIAN

1.1 Introduction

Semiparametric models such as Cox's (1972) [10] proportional hazards model and linear transformation models (Cheng et al., 1995, 1997 ; Fine et al., 1998) [7] [8] [16] and their special cases (e.g., accelerated failure time model) are very popular for modeling effects of covariates on survival response. For example, the main aim of a semiparametric model for a two-arm randomized trial for small cell lung-cancer (SCLC) patients (Ying et al., 1995) [59] is to express the effects of treatment arm and age at entry on time to death (survival time). The Cox model or proportional hazards (PH) model [10] specifies that the hazard function of survival time T has the form

$$\lambda(t; Z) = \lambda_0(t) \exp(\beta' Z), \tag{1.1}$$

given covariates Z. $\lambda_0(t)$ in the (1.1) is an unspecified baseline hazard funtion, and β is the vector of regression parameters. The above model assumes that the covariates have multiplicative effects on the hazard function. The ratio of the hazard function for two individuals with different covariates is constant. For example, if Z_1 is the treatment effect ($Z_1 = 1$ if treatment and $Z_1 = 0$ if placebo) and all other covariates have the same value, then the ratio of hazard function has

$$\frac{\lambda(t; Z = 1)}{\lambda(t; Z = 0)} = \exp(\beta), \tag{1.2}$$

which is the risk of experiencing the event if the individual in the treatment group relative to the risk of experiencing the event if the individual with the placebo group.

Often in practice, there is substantial data information available about the median survival response for a particular trial or study. However, previous semiparametric models (e.g., Cox's (1972) [10])for survival data do not focus on the effects of covariates on the median and other quantiles. Several authors including Ying et al. (1995) [59] gave compelling arguments in favor of focusing on the quantiles of the survival time for modeling and reporting of data analysis results. The effect of treatment and age on the quantiles including median

time to death is useful for describing covariate effects. Clinical trials based on survival outcomes are often designed to detect differences in median survival between treatment arms. Some models based on median are also useful in dealing with heteroscedasticity.

Particularly for Bayesian survival analysis, medians and other quantiles are natural choices for elicitation of experts opinions. Clinical experts on the disease under study are likely to have useful prior information/opinions about survival quantiles (say, the median) and the changes in the median for varying covariate values. However, semiparametric Bayesian models for survival data, possibly with the exception of Kottas and Gelfand (2001), and Hanson & Johnson (2002) [33] [23], are either based on covariate effects on the hazard ratio (see Ibrahim et al., 2001) [26] or on the mean survival time (e.g., Kuo and Mallick, 1997; Walker and Mallick, 1999) [34] [55]. In two-arm cancer clinical trials, the determination of a clinically significant difference and subsequent evaluation of power of the trial, even for frequentist trial designs, are often based on the prior evaluation of the median for the control arm as well as the clinically significant effect of treatment on median survival time (Piantadosi, 2005) [46]. In this paper, our goal is to propose novel semiparametric models for median survival time with interpretable regression effects. We show that these wide classes of semiparametric models have many desirable properties including model identifiability and non-monotone hazards. Unlike previous methods for Bayesian survival analysis (e.g., Hanson & Johnson, 2002) [23], our models accommodate the situation when the location/median as well the scale and shape of the survival distribution are affected by the covariate. Unlike some of the previous frequentist methods for median regression, we do not require the restrictive assumption that all quantile functions below the median to be linear.

In Section 1.2, we introduce two new semiparametric survival models with log-linear median regression functions. In Section 1.2, we also show the desirable properties of these two large classes of survival models, including closed form expressions for other quantile functions (besides the median). We also present the comparisons as well as relationships of our models with existing Bayesian and frequentist median regression models. In Section 1.3 and 1.4, we present the likelihood, suitable nonparametric prior processes and MCMC (Markov Chain Monte Carlo) tools to estimate the model parameters using a semiparametric Bayesian approach. In section 1.5, we consider the SCLC trial to demonstrate how our model can facilitate the determination of prior distributions based on prior opinions about two quantiles (or survival probabilities at two predetermined time-points) of a random patient with known age and treatment arm. In section 1.5, we also illustrate the practical utility of our Bayesian methods and model diagnostics via analyzing the SCLC study. We also compare the analysis results with analysis based on existing models. We demonstrate that our data analysis methods have the advantages of simplicity, support of the usual justification of the Bayesian paradigm, and ease of computation and implementation. In Section 1.6, our simulation studies reveal that estimators based on our models have better small sample performances and more robustness properties compared to competing methods for median regressions including the estimators of Portnoy (2003) [47]. Some final remarks are in Section 1.7.

1.2 Semiparametric Models

Let T_i be the survival time of subject $i = 1, \ldots, n$ and let $Z_i = (1, Z_{i1}, \ldots, Z_{ip})'$ be the corresponding vector of p time-constant covariates along with the intercept term. The commonly so called linear transformation model (Cheng et al., 1995, 1997) [7] [8] assumes that

$$h(T_i) = \gamma' Z_i + e_i \ , \tag{1.3}$$

where h is a monotone transformation, $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_p)$ is a vector of regression parameter, and e_i is an unspecified error random variable with common density $f_e(\cdot)$ free of covariate Z_i. Usually the density $f_e(\cdot)$ of e_i is assumed to be a member of some parametric family with location 0 and with shape and scale free of Z_i.

Important special cases of (1.3) are the accelerated failure time model (AFT) when $h = \log$, the proportional odds model when e_i comes from standard logistic distribution with cumulative distribution function (CDF) $F(x) = e^x/(1 + e^x)$, and Cox's model (1972) [10] when f_e is the extreme-value density with CDF $F(x) = 1 - \exp[-\exp(x)]$. It can be seen that one of the most attractive feature of the linear transformation model is its generality as F can be any distribution function. More detailed references that discuss the application of model (1.3) to regression analysis of right-censored survival data include Chen et al. (2002) [6], Cheng et al. (1995, 1997) [7] [8], Fine et al. (1998) [16], Kong et al. (2004) [32], and Lu and Ying (2004) [35].

The monotone power transformation $g_\lambda(y)$ (Bickel and Doksum, 1981) [3],

$$g_\lambda(y) = \frac{\text{Sgn}(y)\ |y|^\lambda}{\lambda} \text{ for } \lambda > 0 \ , \tag{1.4}$$

where $\text{Sgn}(y) = -1$ for $y < 0$ and $\text{Sgn}(y) = +1$ otherwise, is an extension of the Box-Cox power family (Box and Cox, 1964), a popular transformation to obtain symmetric and unimodal density for the transformed random variable. We assume that for unknown λ, the transformed survival time $g_\lambda\{\log(T_i)\}$ is symmetric and unimodal with median $g_\lambda(\beta' Z_i) = g_\lambda(M_i)$, that is,

$$g_\lambda\{\log(T_i)\} = g_\lambda(M_i) + \epsilon_i \tag{1.5}$$

where ϵ_i are iid from a unimodal and symmetric density $f_\epsilon(\cdot)$ centered at 0, $M_i = \beta' Z_i$, and β is the vector of regression parameters. Carroll and Ruppert (1984) [5], Fitzmaurice et al. (2007) [19], among others proposed parametric version of the transform-both-sides (TBS) regression model for an uncensored continuous response with the original Box-Cox transformation (Box and Cox, 1964) [4] and $N(0, \sigma^2)$ density for error $f_\epsilon(\cdot)$.

The transformation $g_\lambda(y)$ in (1.4) is monotone with derivative $g'_\lambda(y) = |y|^{\lambda-1}$. The median of $\log(T_i)$ is $M_i = \beta' Z_i$ because

$$P[\log(T_i) > M_i] = P\left[g_\lambda\{\log(T_i)\} > g_\lambda(M_i)\right] = F_\epsilon(0) = 1/2 \tag{1.6}$$

where F_ϵ is the cdf of ϵ. As a consequence, the survival time T_i has a log-linear median regression function $Q_{0.5}(Z_i) = \exp(M_i) = \exp(\beta' Z_i)$ and survival function $S(t|z) = 1 - F_\epsilon(g_\lambda(\log t) - g_\lambda(M))$. For the small cell lung-cancer (SCLC) study with $M_i = \beta_0 + \beta_1 z_1 + \beta_2 z_2$, where z_1 is a treatment indicator and z_2 denotes age, this implies that the ratio of medians from two patients of the same age but different treatment arms is $Q_{0.5}(z_1 =$

$1, z_2)/Q_{0.5}(z_1 = 0, z_2) = \exp(\beta_1)$. We also get a similar straightforward interpretation of $\exp(\beta_2)$ as the ratio of the medians for unit increase in age. The following theorem shows that the parameter λ and the density f_ϵ of (1.5) are also identifiable, in the sense that for any survival time following (1.5), there is a unique (λ, f_ϵ) for which $g_\lambda\{\log(T_i)\}$ has a symmetric unimodal distribution.

Theorem 1. *For the model in (1.5) if there is another triplet* $(\lambda^*, \beta^*, f_{\epsilon^*})$ *for which* $g_{\lambda^*}\{\log(T)\} = g_{\lambda^*}(\beta^* x) + \epsilon^*$, *then* $\lambda = \lambda^*$, $\beta = \beta^*$ *and* $f_\epsilon = f_{\epsilon^*}$.

The proof of Theorem 1 is in the Appendix A. Similar to the transformation model of (1.3), we can rewrite the TBS model of (1.5) as

$$\log(T_i) = M_i + e_i, \tag{1.7}$$

where the error e_i in (1.7) has asymmetric density function

$$f_e(u|Z_i) = \ f_\epsilon\{g_\lambda(M_i + u) - g_\lambda(M_i)\}\ g'_\lambda(M_i + u), \tag{1.8}$$

where $g'_\lambda(y) = \ |y|^{\lambda-1}$. The shape and scale of the cdf $F_\epsilon\{g_\lambda(M_i+u)-g_\lambda(M_i)\}$ of e_i depend on Z_i. The approximate variance of $\log T$ is $\sigma^2_\epsilon|M|^{2(1-\lambda)}$, where f_ϵ has finite variance σ^2_ϵ. It is clear that unlike the usual assumption of the transformation model of (1.3) and Bayesian models of, say, Hanson & Johnson (2002) [23], the median as well as the shape and scale of the error density $f_e(\cdot|Z_i)$ in (1.7) depend on the covariate Z_i. This allows our model to be useful for dealing with heteroscedasticity of $\log T$. Unlike the existing Bayes models, the covariate Z does affect the scale and shape of the f_e in our TBS models. A parametric log-normal model with location $M(Z) = \beta' Z$ for $\log(T)$ is a special case of (1.5) with $\lambda = 1$ and F_ϵ being $N(0, \sigma^2)$. The hazard function $h(t|Z) = -\frac{d}{dt}\log\{P(T > t|Z)\}$ of (1.5) can be non-monotone; for example, a log-normal model has non-monotone hazard.

In relation to the parametric TBS model with Gaussian ϵ, Kettl (1991) [29] introduced another parametric model by using a Taylor expansion of the transform both sides model, which is written as

$$g_\lambda\{\log(T_i)\} = g_\lambda(M_i) + |M_i|^\gamma \eta_i \tag{1.9}$$

with $\eta_i \sim N(0, \sigma^2)$. As an alternative to the semiparametric TBS model of (1.5), we can consider a semiparametric extension of Kettl's (1991) [29] model with a symmetric unimodal density for η_i. However, for sake of parsimony, we recommend not using two separate parameters λ and γ to achieve symmetry and to model heteroscedasticity respectively. Instead, we suggest the Power-Of-Mean-Scale (POMS) semiparametric model

$$\log(T_i) = M_i + |M_i|^\gamma \epsilon_i \tag{1.10}$$

with $\epsilon_i \sim f_\epsilon(.)$ symmetric and unimodal at 0, as an alternative to the TBS model. The key assumption of the semiparametric POMS model of (1.10) is that $\log(T)$ is symmetric and unimodal with median $M_i = \beta Z_i$. Unlike the Bayes model of Kottas and Gelfand (2001) and Hanson & Johnson (2002) [33] [23], the POMS model of (1.10) takes care of the heteroscedasticity of $\log(T_i)$. The survival function of (1.10) is $S(t|Z) = S_\epsilon\{\log(t^{|M|^\gamma} e^{-sgn(M)|M|^{1-\gamma}})\}$, and it reduces to the accelerated failure time model with $S(t|Z) = S_\epsilon\{\log(te^{-M})\}$ when $\gamma = 0$.

Although the models in (1.5) and (1.10) apparently focus on modeling the median, we can easily obtain other quantiles of $\log(T)$. For TBS of (1.5), the α-quantile $Q_\alpha(Z)$ of T is

$$Q_\alpha(Z) = \exp\{M^*_\alpha(Z)\} = \exp\left[g_\lambda^{-1}\{g_\lambda(\beta' Z) + \epsilon^*_\alpha\}\right] \ , \tag{1.11}$$

because $P[g_\lambda\{\log(T)\} < g_\lambda(M) + \epsilon^*_\alpha \ |Z] = \alpha$ for $\alpha \in (0,1)$, where ϵ^*_α is the α-quantile of $f_\epsilon(\cdot)$ with $P(\epsilon < \epsilon^*_\alpha) = \alpha$. For $\alpha = 0.5$, we have $\epsilon^*_{0.5} = 0$ and get the log-linear median function $\exp(\beta' Z)$ for T in (1.5). Similarly, the α-quantile of T for the POMS model is

$$Q_\alpha(Z) = \exp(M + |M|^\gamma \epsilon^*_\alpha), \tag{1.12}$$

where $M = \beta' Z$ for the regression parameter β in (1.10). These expressions of (1.11) and (1.12) show that both of these models are very convenient for simultaneously estimating all important quantiles of T_i using the estimates of parameters and ϵ^*_α. However, unlike the existing methods including those of Portnoy (2003) [47] and Peng and Huang (2008) [44], $Q_\alpha(Z)$ of the TBS and POMS models in (1.11) and (1.12) are not linear in covariate Z unless $\alpha = 0.5$ (median). The Bayesian models of Kottas and Gelfand (2001) and Hanson & Johnson (2002) [33] [23] also have linear quantile functions $M_\alpha(Z) = \beta'_\alpha Z$ of $\log T$ for all $1 > \alpha > 0$, and they are parallel to each other (with only the intercept of β_α different for different $\alpha \in (0,1)$).

1.3 Model Estimation and Inference

Let T_i and C_i be the survival and censoring times, respectively, for $i = 1, \cdots, n$. We observe (t_{i0}, δ_i), where $t_{i0} = T_i \wedge C_i$ is the observed follow-up time and δ_i is the censoring indicator, with $\delta_i = 1$ for $T_i = t_{i0}$ and 0 otherwise. It is assumed that T_i and the random censoring time C_i are conditionally independent given covariate Z_i. Given the observed data vector $(\mathbf{t_0}, \delta^*)$ with $\mathbf{t_0} = (t_{10}, \ldots, t_{n0})$ and $\delta^* = (\delta_1, \ldots, \delta_n)$, the likelihood function under our TBS model of (1.5) is as follows:

$$L(\beta, \lambda, f_\epsilon | \mathbf{t_0}, \delta^*) \propto \prod_{i=1}^n \left\{ |y_i|^{\lambda-1} f_\epsilon(\omega_i) \right\}^{\delta_i} \{1 - F_\epsilon(\omega_i)\}^{1-\delta_i}, \tag{1.13}$$

where $\omega_i = g_\lambda(y_i) - g_\lambda(\beta' Z_i)$ with $y_i = \log(t_{i0})$, $F_\epsilon(\omega) = \int_{-\infty}^{\omega} f_\epsilon(u)du$ is the cdf of the unimodal symmetric density function f_ϵ. For the POMS model of (1.10), the corresponding likelihood is

$$L(\beta, \gamma, f_\epsilon | \mathbf{t_0}, \delta^*) \propto \prod_{i=1}^n \left\{ f_\epsilon(\omega_i^*) \, |M_i|^{-\gamma} \right\}^{\delta_i} \left\{ \overline{F}_\epsilon(\omega_i^*) \right\}^{1-\delta_i}, \tag{1.14}$$

where $\omega_i^* = (y_i - M_i)/|M_i|^\gamma$, $M_i = \beta' Z_i$ and $\overline{F}_\epsilon(u) = \int_u^{+\infty} f_\epsilon(z)dz$.

In general, for the parametric versions of either TBS or POMS model, any unimodal symmetric distribution, such as the Gaussian and logistic, can be used for F_ϵ. For example, $f_\epsilon(w)$ and $F_\epsilon(w)$ will be respectively replaced by the density $\phi_\sigma(w)$ and cdf $\Phi_\sigma(w)$ of $N(0, \sigma^2)$ for the Gaussian TBS model likelihood in (1.13) and the Gaussian POMS model's likelihood in (1.14). The corresponding posterior is $p(\tau, \sigma | \mathbf{t_0}, \delta^*) \propto L(\tau, \sigma | \mathbf{t_0}, \delta^*)\pi(\tau, \sigma)$, where $\pi(\tau, \sigma)$ is the joint prior density based on the available prior information, with $\tau = (\beta, \lambda)$ for

TBS and $\tau = (\beta, \sigma)$ for POMS. Markov Chain Monte Carlo (MCMC) samples from this joint posterior can be used to implement a parametric Bayesian analysis. Under either of these two parametric models, the maximum likelihood estimator (MLE) of the regression parameters β can be obtained via maximizing the log-likelihood $L(\tau, \sigma|\mathbf{t_0}, \delta^*)$ corresponding to the chosen model. For example, the log-likelihood function of the (Gaussian) parametric TBS model is

$$\ell(\beta, \lambda, \sigma|\mathbf{t_0}, \delta^*) = \sum_{i=1}^{n} \left\{ \delta_i \log \phi_\sigma(\omega_i) + \delta_i(\lambda - 1) \log(|y_i|) + (1 - \delta_i) \log \overline{\Phi}_\sigma(\omega_i) \right\} \ , \quad (1.15)$$

where $\overline{\Phi}_\sigma(\omega) = 1 - \Phi_\sigma(\omega)$ is the survival function of $N(0, \sigma^2)$. For the sake of brevity, we skip the log-likelihood expression for the POMS model of (1.10). The maximum likelihood estimator (MLE) of the parameters under either parametric TBS or POMS models is obtained via maximizing the corresponding log-likelihood function $\ell(\beta, \tau|\mathbf{t_0}, \delta^*)$ using Newton-Raphson (NR) iterations. Under mild regularity conditions, the MLE of β (as well as the parametric Bayes estimator) is consistent and asymptotically efficient based on regular large sample theory for the MLE when the modeling assumption is correct.

1.4 Semiparametric Bayesian

Any parametric assumption about f_ϵ is deemed as a restrictive parametric assumption for some data examples in practice. In the semiparametric models of (1.5) and (1.10), the unimodal symmetric distribution F_ϵ of error ε_i is assumed unknown. The semiparametric likelihood of model in (1.5) is given as

$$L(\beta, \lambda, F_\epsilon|\mathbf{t_0}, \delta^*) \propto \prod_{i=1}^{n} \left\{ |y_i|^{\lambda - 1} \ dF_\epsilon(\omega_i) \right\}^{\delta_i} \left[\bar{F}_\epsilon(\omega_i) \right]^{1 - \delta_i} \ , \quad (1.16)$$

where $\bar{F}_\epsilon(u) = 1 - F_\epsilon(u)$ and $\omega_i = g_\lambda(y_i) - g_\lambda(\mu_i)$. Similarly, the semiparametric likelihood of POMS model of (1.10) is

$$L(\beta, \gamma, F_\epsilon|\mathbf{t_0}, \delta^*) \propto \prod_{i=1}^{n} \left\{ |M_i|^{-\gamma} \ dF_\epsilon(\omega_i^*) \right\}^{\delta_i} \left[\bar{F}_\epsilon(\omega_i^*) \right]^{1 - \delta_i} \ , \quad (1.17)$$

where $\omega_i^* = (y_i - M_i)/|M_i|^\gamma$ for $M_i = \beta' Z_i$. For semiparametric maximum likelihood estimation (SPMLE) under these models, the likelihoods are maximized with respect to the restriction that F_ϵ is the cdf of a unimodal distribution symmetric around 0. The regularity conditions and asymptotic issues for the SPMLE under either (2.9) or (1.17) are nontrivial and beyond the scope of this paper. For semiparametric Bayesian analysis, we need the posterior

$$p(\tau, F_\epsilon|\mathbf{t_0}, \delta^*) \propto L(\tau, F_\epsilon|\mathbf{t_0}, \delta^*) \pi_{12}(\tau) \pi_3(F_\epsilon) \ , \quad (1.18)$$

where π_{12} and π_3 are independent priors of τ and F_ϵ. We would like to emphasize that the expression of the likelihood $L(\tau, F_\epsilon|\mathbf{t_0}, \delta^*)$ in (1.18) and even the priors π_{12} and π_3 actually depend on the underlying semiparametric model (either TBS of (1.5) or POMS of (1.10)). This uses the simplifying, however reasonable, assumption that the prior opinions about

parametric vector τ and nonparametric function F_ϵ can be specified independently. We will discuss the practical justification of this assumption later.

We now introduce a class of nonparametric priors π_3 for F_ϵ, applicable either for model in (1.5) or in (1.10), defined over the space of symmetric unimodal distribution functions. We use the result that any symmetric unimodal distribution F_ϵ can be expressed as a scale-mixture of uniform random variables

$$F_\epsilon(u) = \int_0^\infty \zeta(u|\theta)\ dG(\theta) \tag{1.19}$$

for some mixing distribution $G(\theta)$ (Feller, 1971, p.158) [14], where $\zeta(u|\theta)$ for $\theta > 0$ is the uniform distribution with mean zero and support $(-\theta, +\theta)$. We use the Dirichlet process (Ferguson 1973) [15] $G \sim DP(G_0, \nu)$ prior for the unknown scale-mixing distribution $G(\theta)$ of (1.19) to define a nonparametric prior for the random (unknown) unimodal symmetric distribution F_ϵ. The Dirichlet process (DP) is characterized by the known "prior guess" G_0 (the prior expectation of G), and a positive scalar parameter ν, precision around prior mean/guess G_0. For a measurable set B on the real line, the Dirichlet process prior assumes that for any partition $(B_1, \ldots, B_m)$, the joint distribution of the random vector $(G(B_1), \ldots, G(B_m))$ is given by the Dirichlet $(\nu G_0(B_1), \ldots, \nu G_0(B_m))$.

The prior mean G_0 of the random mixing density G can be chosen appropriately to assure a desired prior mean/guess F_* for unknown F_ϵ. Using a result by Khintchine (1938) [30], when the density $f_*(\cdot)$ and its derivative $f'_*(\cdot)$ exist, the density $G'_0(\theta)$ of $G(\theta)$ is given as

$$G'_0(\theta) = -2\theta\ f'_*(\theta) \quad \text{for } \theta > 0\ . \tag{1.20}$$

For example, to obtain an approximate double exponential ($Dexpo(\gamma)$) prior mean density $f_*(\epsilon) = \frac{1}{2}\gamma \exp(-\gamma|\epsilon|)$ for error density f_ϵ, using (1.20), we need to choose $G_0(\theta|\gamma)$ as $Gamma(2, \gamma)$ with density $G'_0(\theta|\gamma) = \gamma^2\theta \exp(-\gamma\theta)$. The precision parameter ν also determines the degree of belief about how close F_ϵ should be to its prior guess F_*. When ν is large enough, the unknown nonparametric F_ϵ is very close to its prior mean/guess $F_*(\cdot|\gamma)$ (which is pre-specified and often parametric). When ν is small, we have very little confidence in unknown F_ϵ being close to $F_*(\cdot|\gamma)$, and the corresponding Bayes estimators of the regression parameters are expected to be very close to the semiparametric likelihood estimator of the corresponding semiparametric model. The details of the specifications of the hyperparameters of the priors π_{12} and π_3 in (1.18) are provided in the data analysis section.

1.5 Data Example

We analyze the data set from the randomized cross-over trial of Etoposide (E) and Cisplatin (C) for small cell lung cancer (SCLC) patients (Ying et al., 1995) [59]; 62 cancer patients ($z_1 = 1$) are randomized to arm A (C followed by E) and 59 patients ($z_1 = 0$) to arm B (E followed by C) (Figure 1.1). Apart from treatment indicator z_1, another covariate is the patient's age at entry (z_2) centered at age 50. Each survival time (in months) was either observed ($\delta_i = 1$) or administratively censored ($\delta_i = 0$). To evaluate the age-adjusted treatment difference, we consider the linear regression function $M_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i}$.

The maximum likelihood estimates of the regression parameters β under parametric TBS model (1.5) with Gaussian F_ϵ are given by $\hat{\beta}_0 = 3.349$, $\hat{\beta}_1 = 0.433$, $\hat{\beta}_2 = -0.019$.

Now we present a parametric Bayesian analysis using the TBS model of (1.5) with parametric $N(0, \sigma^2)$ density for F_ϵ. One major advantage of the TBS model for Bayesian analysis is that the priors for the parameters $(\beta_0, \beta_1, \beta_2, \lambda, \sigma)$ can be determined based on prior opinions about some key quantities related to the prior-predictive survival time T^* of a patient with known covariate values, say, (z_1^*, z_2^*). Without loss of generality, we take $z_1^* = 0$ and $z_2^* = 0$, that is the priors are based on the following: (1) Prior guess and prior range of a quantile, say, median, of the prior-predictive survival time T^* of a patient of age 50 from the control arm (treatment B); (2) Change in median survival of that patient for unit change in each of age (z_2) and treatment (z_1). We point out that for most Phase 2 and 3 trials, these quantities are routinely elicited and used to design the trial and determine the power for detecting differences (e.g., Pintadosi, 1997) [46]. We first demonstrate how the priors for the parametric TBS models can be determined. The prior $\pi(\beta, \gamma, \sigma)$ for the parametric POMS model can be determined in a similar way, because of the relationship between two models when $1 - \lambda = \gamma$.

We use the simplifying assumption that the joint prior is $\pi(\beta, \lambda, \sigma) = \pi_1(\beta)\pi_2(\lambda)\pi_3(\sigma|\beta, \lambda)$. This assumption can be justified in practice because prior $\pi_1(\beta)$ is based on median (location) of T^*, whereas the prior $\pi_2(\lambda)$ is based on the shape (skewness) of $\log(T^*)$. The specification of the prior for β_0 uses the fact that T^* with $z_1^* = z_2^* = 0$ has a prior median $\exp(\beta_0)$. For the lung cancer trial conducted before 1993, the current expert opinions about SCLC are not very appropriate. Based on the published literature about the treatment of SCLC before this study (e.g. Jett J.R. et al., 1990; Evans W.K. et al., 1987; Comis R.L., 1986) [27],[12], [9], the median survival time for treatment arm B was thought to be between 12 to 17 months for limited-state patients and 9 to 10 months for extensive-stage SCLC patients. For our SCLC study with nearly equal proportions of these two types of SCLC patients, we use a mean prior guess of 13 months and a range of (8, 18) months for T^*. These give us the prior $\beta_0 \sim N(A_1, B_1^2)$ with $A_1 = \log(13)$ and $B_1 = \{\log(18) - \log(8)\}/3$ to ensure that the prior range of β_0 has approximate length $3B_1$. Our prior opinion about β_1 is based on the prior belief about the ratio of medians $\{Q_{0.5}(z_1 = 1, z_2^*)/Q_{0.5}(z_1 = 0, z_2^*)\} = \exp(\beta_1)$ of two patients with identical age, but, from different treatment arms. So, the prior $\beta_1 \sim N(0, 10)$ corresponds to a 95% prior probability that the ratio of medians e^{β_1} is in $(e^{-2\sqrt{10}}, e^{2\sqrt{10}})$ and centered at $e^0 = 1$ (indifferent opinion regarding superiority of either treatment arm). Similarly, the prior $\beta_2 \sim N(0, 10)$ corresponds to prior opinion that two patients from same treatment and with 1 year difference in age, have a ratio of medians between $(e^{-\sqrt{10}}, e^{\sqrt{10}})$ with 68% probability. We have chosen such a non-informative prior opinion about β_1 and β_2 to allow for a meaningful comparison of our analysis results with results from frequentist and previous Bayes methods based on either no prior or non-informative prior. We would like to point out that our point-wise Bayes estimates do not change substantially ($< 4\%$ change) when we reduce the prior variances of β_1 and β_2 to 1 (instead of 10). The interval estimates of β_1 (as an example) is around 12% narrower when using these more skeptical (of regression effects) $N(0, 1)$ priors instead of using $N(0, 10)$ priors for β_1 and β_2.

We use the $Unif(0, 3)$ prior for $\pi_2(\lambda)$ because it is difficult to interpret the after-transform linear model of (1.5) when $\lambda > 3$. In their original paper, Box and Cox (1964)

recommended restricting the $\lambda \leq 2$. For parametric Gaussian TBS model, $\log T^*$, when $z_1^* = z_2^* = 0$, can be expressed approximately as $\log T^* \simeq \beta_0 + \sigma|\beta_0|^{1-\lambda}e^*$ (Kettl, 1991) [29], where β_0 is the median of $\log T^*$ and $e^* \sim N(0,1)$. This allows us to obtain prior $\pi_3(\sigma|\beta_0, \lambda)$ based on prior opinion of $M^*_{\alpha^*}$ because

$$|\beta_0 - M^*_{\alpha^*}| \simeq \sigma|\beta_0|^{1-\lambda}|e^*_{\alpha^*}| \Rightarrow \sigma \simeq \frac{|\beta_0 - M^*_{\alpha^*}|}{|\beta_0|^{1-\lambda}|e^*_{\alpha^*}|}, \tag{1.21}$$

where $M^*_{\alpha^*}$ is another quantile of $\log T^*$ for $\alpha^* \neq 1/2$, and $e^*_{\alpha^*}$ is the α^*-percentile of standard normal. For example, when we take $\alpha^* = 0.75$, we have $\sigma \simeq |\beta_0 - M^*_{0.75}|\ |\beta_0|^{\lambda-1}/0.6745$. Based on SCLC literature prior to this trial, we use the prior opinion that the third-quartile $\exp(M^*_{0.75})$ of a patient in treatment arm with 50 years entry-age is between 10 months to 5 years with center at 33 months. For given (β_0, λ), we use a Gamma density at the prior $\pi_3(\sigma|\beta_0, \lambda)$ with mean equal to $|\beta_0 - \log(33)||\beta_0|^{\lambda-1}/0.6745$ and approximate range between 0 and to $(\log(60) - \log(10))|\beta_0|^{\lambda-1}/0.6745$. These prior densities give us approximately the same means and ranges of $M^*_{0.5} = \beta_0$ and $|\beta_0 - M^*_{0.75}|$ we expect from our prior opinion about these two quantiles of $\log(T^*)$. However, to simplify this further, we use an unconditional Gamma prior $\pi_3(\sigma)$ whose mean equals to $\frac{|\log(13)-\log(33)|}{0.6745}$ and variance equals to $\frac{|\log(13)-\log(60)|}{0.6745}$ (based on prior mean $\log(13)$ for β_0 and prior guess 1 for λ). We found no noticeable difference in posterior estimates using this unconditional prior for σ instead of a conditional prior $\pi_3(\sigma|\beta_0, \lambda)$. We remind the reader that the priors used in our analysis are solely for demonstrating the method of development of one set of priors for the Bayesian analysis of the lung-cancer study. An expert's prior opinions on median survival time of small cell lung cancer can be very different from what we used, and that may lead to different prior specification of the parameters.

Our plot (left-hand panel of Figure 1.2) of residuals $y_i - y_{i*}$ versus the patient's age at entry, where y_i is the observed log-survival time (subject to censoring) and $y_{i*} = E[\log T_i|\mathbf{t_0}, \delta^*]$ is the posterior predictive expectation of $\log(T)$ under the model, does not show any trend of residuals under parametric Bayes TBS model. Our plot (right-hand panel of Figure 1.2) of these residuals versus the estimated median survival times also does not reveal any serious inadequacy of the parametric TBS model. However, the Q-Q plot (Figure 1.3) of these residuals suggests that the assumption of Gaussian distribution for F_ϵ in (1.5) is questionable due to the plot being non-linear at the right tail. Later, we use the semiparametric Bayesian analyses to avoid the Gaussian assumption of ε_i. Our posterior means (Bayes estimates) of three quartiles $Q_\alpha(z_1, z_2)$ for $\alpha = 0.25, 0.50, 0.75$ of treatment A ($z_1 = 1$) are higher than the corresponding estimated quantiles of treatment B ($z_1 = 0$) at any age ($z_2 > 0$).

For the semiparametric Bayesian analysis with a symmetric unimodal f_ϵ in (1.5), we need to specify the prior guess/mean of F^* of F_ϵ and a prior precision parameter ν. We take the precision parameter $\nu = 1$ to imply a very low confidence around our parametric prior guess F_* of the nonparametric error distribution F_ϵ. We take the prior mean f_* of nonparametric f_ϵ to be $N\{0, (\sigma_0)^2\}$ where $\sigma_0 = \frac{|\log(60)-\log(10)|}{0.6745}$, the approximate prior mean of f_ϵ for the parametric Bayes analysis of the TBS model. Using (1.20), this $N\{0, (\sigma_0)^2\}$ density for f_* corresponds to a $Gamma(3/2,\ 1/\{2(\sigma_0)^2\})$ for G_0 in (1.20). The constructive

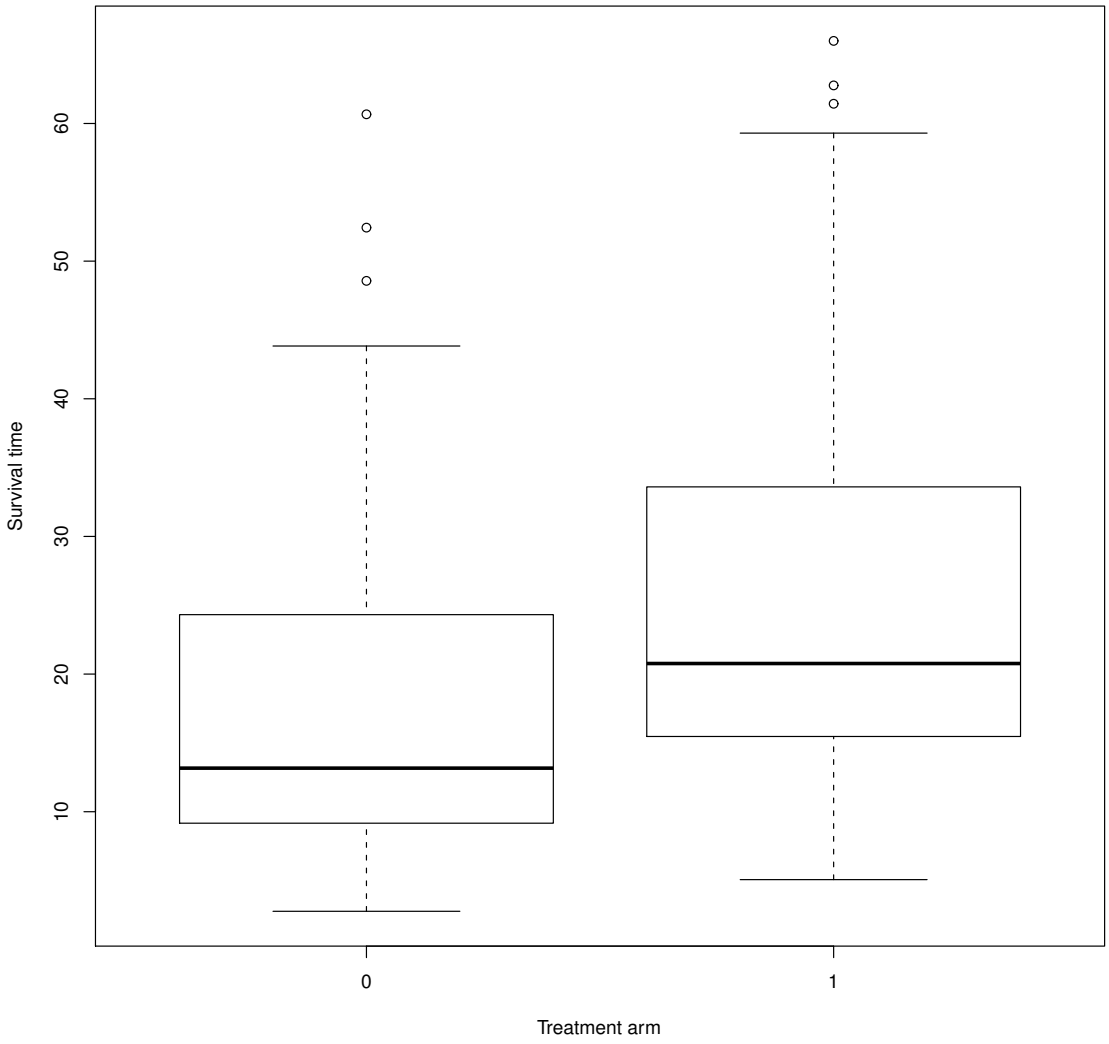

Figure 1.1: Plots of survival time (in months) versus two treatment arms for the lung cancer data. (0 from treatment B, 1 from treatment A.)

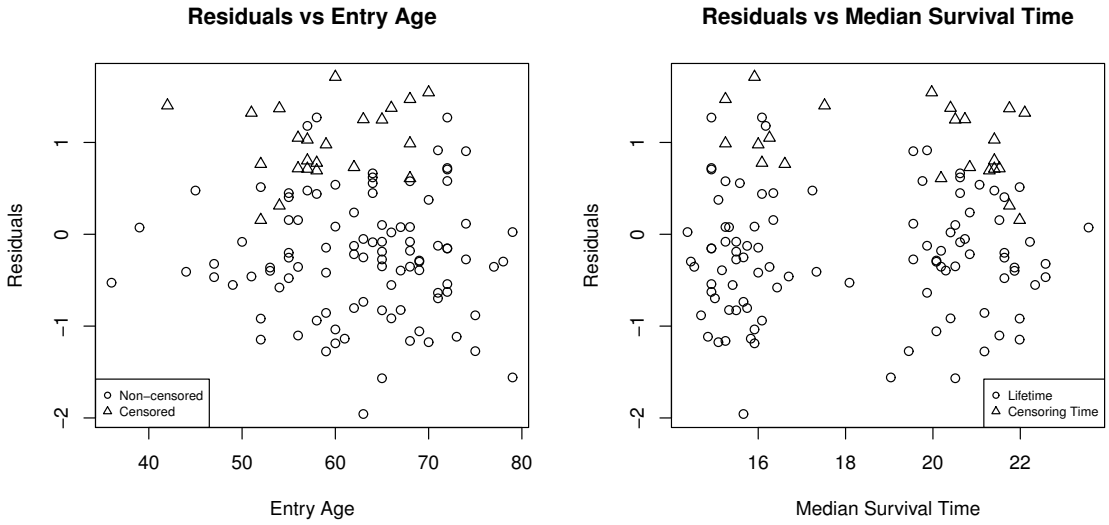

Figure 1.2: Plots of residuals versus the age at entry (in years) and versus the estimated median survival time (in months) using parametric TBS model for the lung cancer data

definition of the DP mixture prior process for F_ϵ is (Sethuraman, 1994) [49]

$$F_\epsilon(u) = \sum_{k=1}^{\infty} p_k \zeta(u|\theta_k), \tag{1.22}$$

where $\theta_k \overset{i.i.d.}{\sim} G_0$, $p_k = V_k \prod_{j<k}(1-V_j)$ with $V_j \overset{i.i.d.}{\sim} Beta(1,\nu)$. The actual implementation of the MCMC tool to sample from (1.18) is based on a finite approximation $F_\epsilon(u) \simeq \sum_{k=1}^{N} p_k \phi(u|\theta_k)$ of (1.22). The MCMC computational tool can be implemented, even in a standard package such as Winbugs, using a finite (say, $N = 1,000$) number (maximum number of θ_k in (1.22)) of $V_1, \ldots, V_N$ and $p_1, \ldots, p_N$. The priors for $V_1, \ldots, V_{N-1}$ are i.i.d. $Beta(1,\nu)$ and $V_N = 1$. After simulating $V_1, \ldots, V_N$, we obtain $p_1 = V_1$ and $p_k = V_k(1-V_{k-1})\ldots(1-V_1)$ for $k = 2, \ldots, N$. The rest of the conditional posteriors are the same as those used for the parametric Bayes.

We get the semiparametric Bayes point estimates $\hat{\beta}_0 = 3.086$, $\hat{\beta}_1 = 0.304$ and $\hat{\beta}_2 = -0.006$ of the regression parameter $(\beta_0, \beta_1, \beta_2)$ along with 95% credible intervals $(2.836, 3.315)$, $(0.003, 0.577)$ and $(-0.022, 0.011)$ (respectively). The results of the Bayes estimators of regression parameters (β_1 and β_2) under parametric and semiparametric TBS models along with ML estimator based on parametric Gaussian error TBS model are presented in Table 1.1. For comparison with TBS models based analysis, we also present parametric and semiparametric analysis of the Power-Of-Mean-Scale (POMS) model of (1.10). We again emphasize the similarities between the interpretation of β and median $M = \beta' Z$ in TBS and POMS models, and their common rule of accommodating heteroscedasticity of $\log T$ with $\lambda = 1 - \gamma$. The only major difference is that the POMS model assumes $\log T$ to be sym-

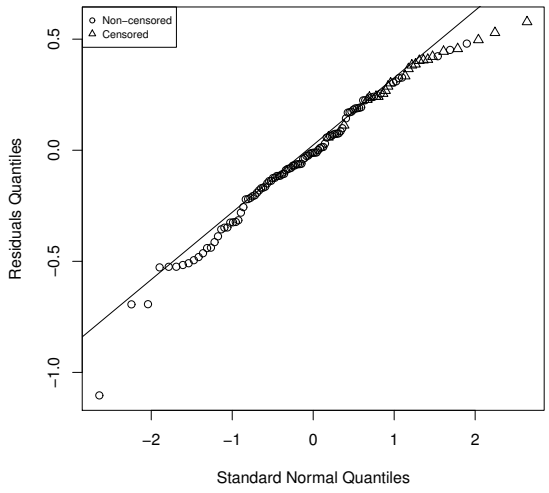

Figure 1.3: Q-Q plots of the residuals under parametric TBS model for the lung cancer data

metric and unimodal, without any need for further transformation. For the Gaussian error POMS model, we use the prior $\pi(\beta, \sigma, \gamma)$ with assumption $\pi(\beta, \sigma, \gamma) = \pi_1(\beta)\pi_2(\gamma)\pi_3(\sigma|\beta, \gamma)$ and matching π_1, π_2 and π_3 with corresponding priors in the TBS model with $\lambda = 1 - \gamma$. For the prior process for f_ϵ in the semiparametric POMS model, we use the same prior process used for f_ϵ in the TBS model. The last line of Table 1.1 is the result for the Bayesian median regression model of Kottas and Gelfand (2001) using the model of (1.7) with $f_e(u) = (1/2)|\eta|^{sgn(u)} f_0(|\eta|^{sgn(u)}u)$ for a nonparametric density $f_0(u)$ defined on $u > 0$.

Table 1.1: Pointwise and 95% interval estimates (within parenthesis) of regression parameters (β_1 for treatment z_1 and β_2 for age z_2) for the lung cancer study under different procedures

Estimator	Treatment	Age
MLE (TBS model)	0.433 (0.141, 0.727)	-0.019 (-0.037, -0.002)
Parametric Bayes (TBS)	0.318 (0.036, 0.604)	-0.008 (-0.023, 0.008)
Semiparametric Bayes (TBS)	0.304 (0.083, 0.577)	-0.009 (-0.021, -0.002)
Parametric Bayes (POMS)	0.267 (0.010, 0.528)	-0.006 (-0.020, 0.008)
Semiparametric POMS Bayes	0.304 (0.032, 0.581)	-0.005 (-0.020, 0.010)
Portnoy	0.369 (0.149, 0.591)	-0.009 (-0.031, 0.012)
KG Bayes	0.389 (0.037, 0.845)	-0.018 (-0.028, -0.007)

The point estimates of the regression parameters of the median functional under different methods are not strikingly different to the corresponding point estimator obtained via Portnoy's method (2003) [47]. This is also evident from Figure 1.4, where 3 estimated quantiles for Portnoy's method (dotted straight lines) and for semiparametric Bayes TBS model (solid curved lines) are plotted (separately for 2 treatment arms).

We find that the proportion of observations in each quantile-interval is closer to the expected proportions for Bayes estimates of quantile functions compared to Portnoy's method (2003) [47]. However ML and Bayes methods yield smaller estimated standard errors and substantially narrower interval estimates than those obtained using Portnoy's method (2003) [47]. For this data example, the estimates based on TBS models have smaller estimated standard errors for the treatment effect compared to competing procedures. The corresponding posterior standard deviations from parametric and semiparametric Bayes are substantially smaller than the standard errors from Portnoy's method (2003) [47] (at least for age-effect) and those from Kottas and Gelfand's (2001) [33] Bayes methods. This is not surprising because Portnoy's median regression methods have a far larger number of regression parameters than the finite dimensional regression parameter β in (1.5). For this particular data example, we notice that the estimates from semiparametric TBS are very close to the semiparametric POMS model. This is not unexpected because the POMS model is based on the Taylor expansion of TBS model when $\log(T)$ is approximately symmetric. Our results suggest that the log-survival time $\log T$ for this study is approximately symmetric. However, $\log T$ has heteroscedasticity which is well accommodated by semiparametric POMS model. In some other examples, this assumption of symmetric $\log T$ may not hold.

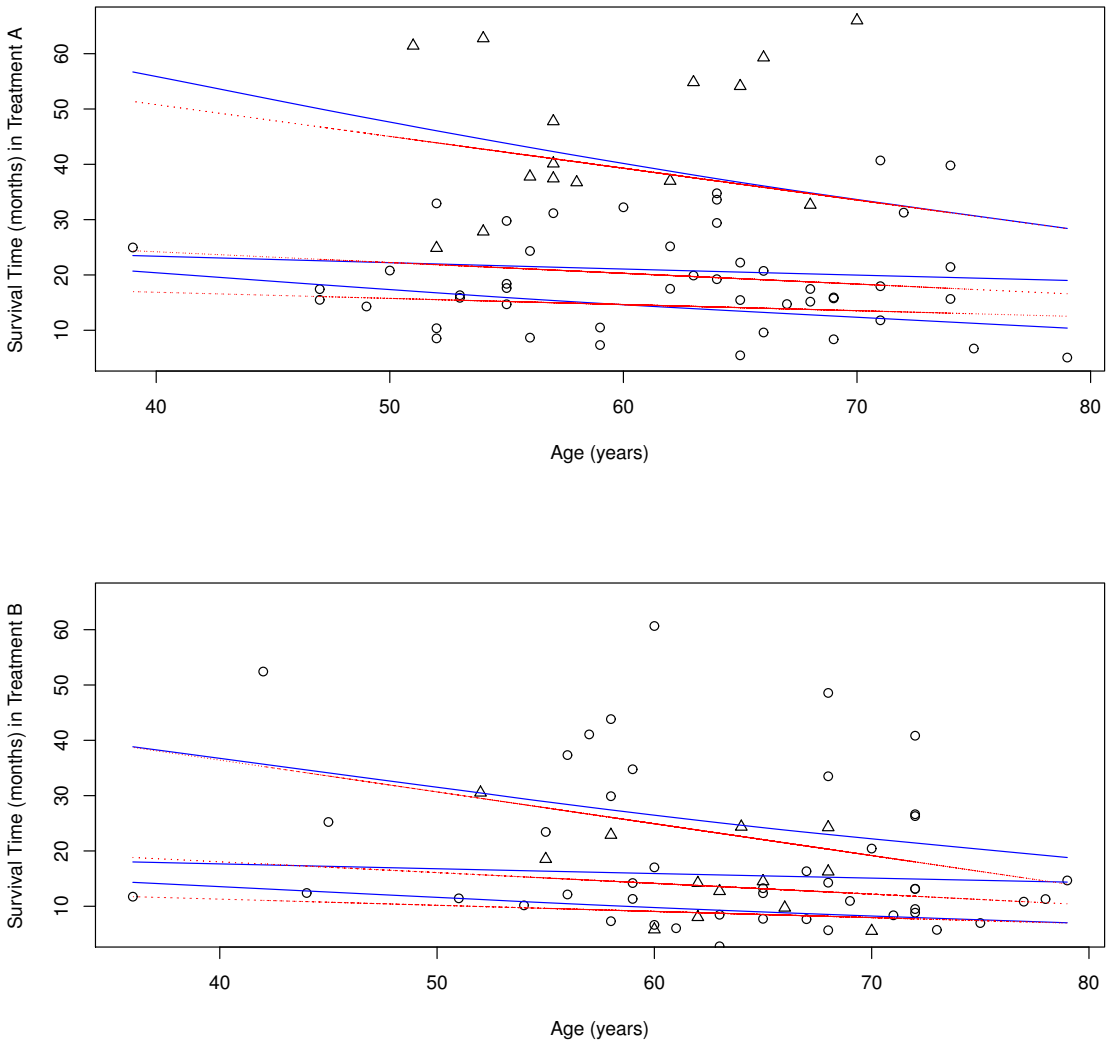

Figure 1.4: Plots of observed survival times versus Age (z_2) with three estimated quartile functions for two treatment arms. (Solid lines: estimated via TBS model; Dotted straight lines: estimated via Portnoy's method; $\triangle$: censored observation)

For Bayesian model comparison, predictive approach has been widely used in the statistical literature. In the data analysis, Conditional Predictive Ordinate (CPO) statistics is calculated for model selection. CPO is a cross-validated probability which gives the marginal posterior predictive density of each point given the rest of the observed data. For references, readers are referred to the paper by Gelfand, Dey, and Chang [20], and the book by Ibrahim, Chen, and Sinha [26]. Figure 1.5 plots the logarithms of CPO (Conditional Predictive Ordinate) ratios for semiparametric TBS model versus POMS model against the observation numbers. A value greater than 0 for logarithms of CPO ratio supports TBS model over POMS model. In this example, only a small majority of observations favor the semiparametric TBS model over the semiparametric POMS model. Our model comparison (Figure 1.6) also shows that for approximately 67% of the observations, the CPO for semiparametric TBS model are higher than the corresponding CPO under parametric TBS model, suggesting an substantially higher proportion of observations supporting the semiparametric model over parametric model. The log CPO plot for parametric POMS versus semiparametric POMS also show similar strong support for the semiparametric model. The final conclusion is that both semiparametric models fit the data better than other competing parametric models, however, there is no clear winner between two competing semiparametric models for this data example.

1.6 Simulation Study

In this section, we use simulation to demonstrate performance of proposed models (1.5) and (1.10). For our simulation setting, we let the median of $Y = \log(T)$ given Z to be $M(Z) = \beta_0 + \beta_1 Z = 6.5 + Z$ with $\beta_0 = 6.5$ and $\beta_1 = 1.0$, where Z can take four possible values $0, 0.5, 1.0$, and 1.5, in equal proportions for each simulated data set. For each simulation distribution of T considered in the study, we simulate at least 5000 datasets with sample sizes $n = 80, 160$, and 320. The number of simulated datasets for different sample sizes may vary to assure that the Monte Carlo variability of the approximate bias and MSE of the regression estimates being smaller than 0.01.

The true median survival time here is set to be $\exp(6.5 + 1 \times 1.5) = 2980.96$ for $z = 1.5$. For the simulation study, the Bayes estimators considered by us are based on semiparametric models of (1.5) and (1.10). The priors used for Bayes estimation in either the TBS or POMS model in the simulation study are: $\beta_0 \sim N(6, 10)$ and $\beta_1 \sim N(0, 1)$. The prior mean for the Dirichlet process is $N(0, 1)$ and the precision is $\nu = 0.01$. This prior for β_1 implies that there is almost 5% prior probability that the ratio of medians is larger than 7.4 for a unit change in z. This prior is very non-informative about the effect of the covariate because it allows a high magnitude in the covariate effect. In order to avoid the undue influence of the choice of the priors on the results of our simulation study, we used somewhat vague priors in the simulations. However, the prior can also be viewed as a skeptical prior because the prior of the regression parameters is centered at the prior guess of no-covariate effect ($\beta = 0$). If a Bayes estimator can demonstrate good performance for detecting covariate-effects with this prior, that suggests that even a skeptical and unusually "flat" prior may not hinder the Bayes method's ability to detect the covariate effect. The implications of chosen priors for multiple model parameters are best described via various summaries of

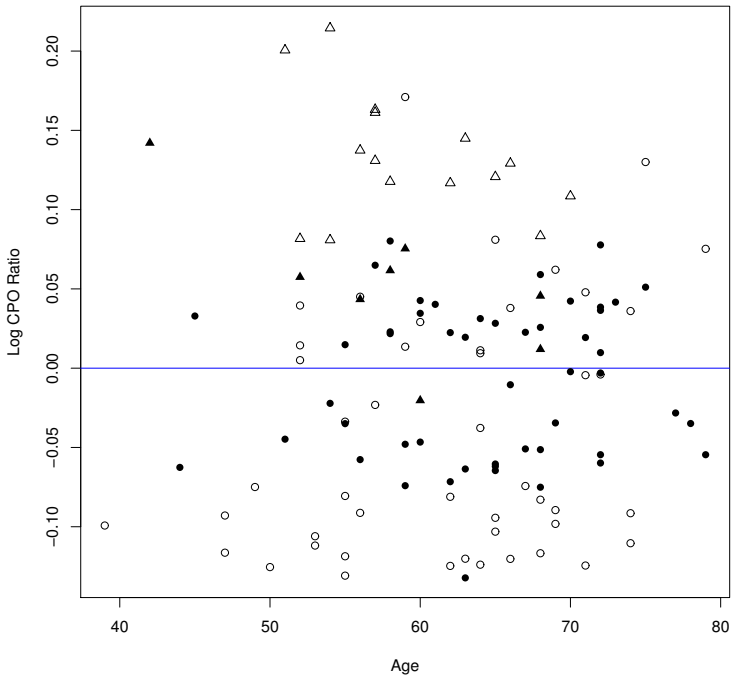

Figure 1.5: Plot of the log-ratio of two CPOs obtained from semiparametric TBS and semiparametric POMS model (y-axis), versus Age (x-axis): ○ uncensored from treatment A; △ censored from treatment A; ● uncensored from treatment B; ▲ censored from treatment B)

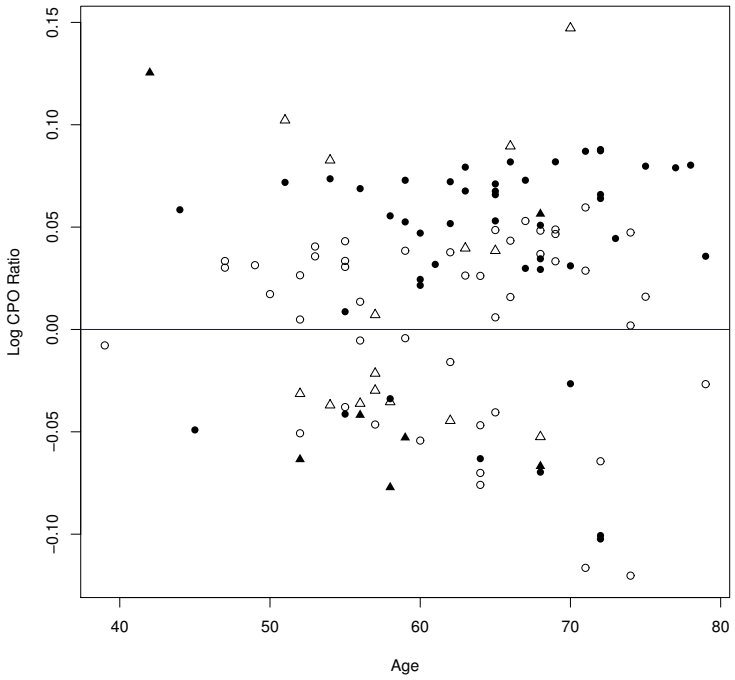

Figure 1.6: Plot of the log-ratio of two CPOs obtained from semiparametric TBS and Gaussian TBS model (y-axis), versus Age (x-axis): ○ uncensored from treatment A; △ censored from treatment A; ● uncensored from treatment B; ▲ censored from treatment B)

the prior predictions of the observables/responses. We generate various summary statistics (e.g. sample mean, range and width of the range) of 500 survival times using a single set of parameters simulated from the joint prior. We then replicate the whole process of simulating these summary statistics 1000 times. We found the range of the sample medians of these 1000 simulated data is between 1400-5200. We can see that our prior predictive models are actually far from being accurate in predicting the true median which is approximately 2981 for $z = 1.5$. The range of survival times from the prior predictive models may have width as large as 10^8. These summaries indicate that our prior predictive models are very non-informative and can cover a wide range of survival patterns. The ranges of our prior predictions are much higher than the true 95% prediction interval under the simulation model. In practice, we expect to use a more informative prior predictive model using often available information about the range of responses (even after incorporating a skeptical prior view about the covariate effect).

First we evaluate the robustness of the maximum likelihood estimators (MLE) and of the semiparametric Bayes estimates based on the TBS model of (1.5) and the POMS model of (1.10). We compare performances (bias and MSE) of these three estimators to the competing frequentist estimator of Portnoy's method (2003) [47]. For this aim, we simulate data from a parametric conditional distribution of Y given Z, namely the Exponential and Pareto, where the assumptions of TBS in (1.5) and POMS in (1.10) are not valid. Both exponential and pareto simulation densities are heteroscedastic and skewed for all λ.

The independent censoring was generated from an Exponential density with rate parameter k chosen to obtain desired proportions of censoring. For example, the choice of $k = \log(2)/30$ results in approximately 20% censoring for exponential simulation model.

Table 1.2 presents the summary of the approximate sampling mean and mean-square-error (MSE) of various competing estimators of β_1 under exponential and Pareto simulation models. Results in Table 1.2 under an exponential simulation model show that the MLE based on (1.5), and the Bayes estimators based on either (1.5) or (1.10) have comparable biases relative to competing estimators. Further, the MSE of Portnoy's estimators are much larger than the corresponding MSE of the MLE and Bayes estimators. The Bayes estimators under (1.5) and (1.10) have much smaller MSE compared to the MLE.

In simulations for the Pareto distribution with scale parameter equal to 1, $g_\lambda(Y)$ is an extremely skewed and heavy-tailed density for all values of λ. As we expect, the bias of the MLE $\hat{\beta}_1$ is slightly higher than that of Portnoy's estimator. The bias of semiparametric Bayes estimators are similar to the bias of Portnoy's estimator. However, they have much smaller MSE than other competing estimators. For the Pareto simulation model, the MSE of Bayes estimators under the TBS model are substantially smaller compared to that of Bayes estimators based on POMS model.

For the last part of Table 1.2, our aim is to investigate the improved performance of the Bayes estimator under a correct TBS modeling assumption compared to rest of the competing estimators. For this aim, we simulated data from TBS model of (1.5) with $\lambda = 0.5$ and double-exponential density for ϵ. We see that Bayes estimators under the TBS models have substantial improvement in MSE compared to competing estimators including the Bayes estimators under the POMS model. The bias of MLE under the Gaussian TBS model increases with sample size. The MSE for Bayesian POMS model is worse (30-45% increase) than that of Bayes TBS model for all sample-sizes.

Table 1.2: Results of simulation study under Exponential and Pareto models: Monte Carlo approximation of the sampling mean and Mean Square Error (MSE) of different estimators of known $\beta_1 = 1$

Simulation		TBS MLE		Portnoy		SP TBS		SP POMS	
Model	Sample	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
	80	0.91	2.66	0.93	4.27	0.92	0.90	1.06	1.50
Exponential	160	0.97	1.35	1.11	2.28	1.08	0.65	1.09	1.12
	320	0.94	0.69	0.96	1.20	0.93	0.48	1.07	0.68
	80	1.00	12.69	1.12	26.98	1.03	0.95	0.73	1.76
Pareto	160	0.99	6.00	1.08	9.96	1.01	0.85	0.86	4.57
	320	0.95	2.75	0.97	4.41	1.02	0.68	0.88	2.23
	80	0.99	1.94	1.01	1.52	1.04	0.72	0.92	0.83
TBS	160	0.96	0.97	0.98	1.69	0.97	0.48	0.92	0.70
	320	0.97	0.51	0.98	1.35	1.03	0.30	0.97	0.72

In summary, when the distribution of $\log T$ after an optimal transformation has a moderate degree of asymmetry, the MLE and Bayes estimators based on either (1.5) or (1.10) have finite sample biases very similar to that of Portnoy (2003)'s [47] estimator. More importantly, the precision of Bayes estimators based on TBS and POMS are better even when the underlying assumptions of either (1.5) or (1.10) are not entirely valid. However, the MLE's performance depends on the degree of symmetry of the distribution of $g_\lambda(Y)$ under optimal λ. The semiparametric Bayes estimators have excellent biases and smallest MSE among all of its competitors. When the modeling assumption of (1.5) is correct, the Bayes estimator based on (1.5) shows much better MSE compared to any competing estimators. This implies that the semiparametric Bayes estimator based on (1.5) is a safer and more robust estimator to use in practice compared to its competitors.

1.7 Discussion

In this paper, we present two new classes of semiparametric models amenable to Bayes estimation of the log-linear median regression function for censored survival data. Similar to previous semiparametric models (e.g., Cox's model) [10], both models have one nonparametric function f_ϵ and a finite dimensional parameter vector. Our Bayes methods have the advantage of the ease of determination of priors and a simple interpretation of the regression parameters via the ratio of median survival times. Our model classes are large since the nonparametric error density f_ϵ in either model has no assumed functional form except the assumption of symmetry and unimodality. We argue that the assumption of unimodality is needed because it justifies the importance of median as the location parameter of interest. As mentioned in previous literature including Box and Cox (1964) [4], the transformation in (1.4) is often an effective tool to obtain symmetry and accommodate heteroscedasticity. Our method can be applied when the covariate Z affects the location as well as the scale and shape of the regression error of $\log T$.

Median regression offers a useful alternative to the popular regression functions of Cox (1972) [10] and the transformation model of (1.3). There is a substantial literature on median regression for censored survival data, including Ying et al. (1995) [59], McKeague et al. (2001) [38] and Bang and Tsiatis (2003) [2]. These methods involve non-linear discontinuous estimating equations that are difficult to solve, often with multiple solutions. The recursive nature of some of these methods (e.g. that of Portnoy (2003) [47]) make the asymptotic justifications and computations complicated. Peng and Huang's (2008) [44] martingale based estimating equations involve minimization of an L_1-type discontinuous convex functions. Unlike estimation with Cox's model (Cox, 1972) [10], martingale based methods may not be the most efficient for estimating regression parameters of the median survival time. For most of these methods, every quantile functional is assumed to be linear in Z, that is $Q_\alpha(Z) = \beta'_\alpha Z$ for all $\alpha \in (0,1)$, where $P\{T > Q_\alpha(Z)\} = \alpha$. This restrictive modeling assumption may not hold true for any real study and very few known stochastic models can satisfy this, except a Gaussian model of $\log T$ with both mean and standard deviation being linear functions of covariate. Unlike our semiparametric models of (1.5) and (1.10), these frequentist linear quantile models have an infinite number of regression parameters β_α for all $\alpha \in (0,1)$. As a consequence, unlike the model of (1.5) and (1.10) there is no simple expression available for survival functions for frequentist quantile regression models. For more in depth discussion about the implementation, comparisons, asymptotic rate of convergence and consequences of the restrictive assumptions for existing quantile regression approaches for censored survival data, we suggest the excellent review by Koenker (2008) [31].

Existing Bayesian median regression models of Kottas and Gelfand (2001) and Hanson & Johnson (2002) [33] [23] have the linear representation of (1.7) with nonparametric asymmetric $f_e(u)$ with median 0 and free of covariate Z. A consequence of this assumption is that all quantile functions of $\log T$ are linear with the same slope (regression coefficient) for each covariate. As we mentioned before, these previous Bayes models cannot accommodate heteroscedasticity of $\log(T_i)$, a very common phenomena in survival data. Most popular competing models such as parametric Weibull and Cox's model (1972) [10] have this property. We believe that our models achieve a sensible compromise between frequentist median regression models and previous Bayesian models via accommodating heteroscedasticity while not restricting to linear functional for all quantiles. Note that, when the F_ϵ either in (1.5) of TBS or in (1.10) for the POMS model is not Gaussian, the parametric MLE $\hat{\beta}$ based on Gaussian ϵ yields a quasi-likelihood estimator of β and the variance of $\hat{\beta}$ can be estimated using the so-called "sandwich" variance estimator (White 1982) [56]. The loss of efficiency of this estimator under a non-Gaussian model is unknown and is beyond the scope of this paper.

In spite of the many similarities in interpretations and forms between the TBS and POMS model, there are some differences between them. The expression in (1.11) for the TBS model also implies that $Q_\alpha(Z_i) \leq Q_\alpha(Z_j) \Leftrightarrow Q_{\alpha'}(Z_i) \leq Q_{\alpha'}(Z_j)$ for all $\alpha, \alpha' \in (0,1)$. This means that under the model in (1.5), ordering between two patients' median survival times implies uniform ordering between their corresponding survival functions over the entire time-axis. This property is similar to Cox's model where ordering between two hazards (as well as survival functions) remain the same over the entire time-axis. However, the POMS model of (1.5) does not satisfy this property. Under the POMS model of (1.10),

survival functions from two treatment groups may cross each other. The key assumption of symmetric $\log T$ in (1.10) may be restrictive in some applications. Our simulation study shows that MSE of estimator under (1.10) is higher than that under (1.5) when $\log T$ has skewed and heteroscedastic density.

Although we focus on modeling the median functional, our methods can be either used via (1.11) or (1.12) to compute even the joint confidence band of any other quantile functional because the quantile functions of (1.11) and (1.12) are know functions of $(\beta, \lambda, \epsilon^*_\alpha)$. For brevity, we have omitted the results of our simulation study showing an excellent accuracy of joint confidence bands of all these quantile functions $(Q_{0.25}(z), Q_{0.5}(z), Q_{0.75}(z))$ under Bayes TBS model of (1.5) (even when the simulation model is Pareto). However, for some diseases, such as cancers with very good prognosis, the main interest may not be on the median, and the goal may be on modeling the quantile $Q_\alpha(Z)$ as a log-linear function with $P\{T < Q_\alpha(Z)\} = \alpha$ for $\alpha > 1/2$. In this case, we can use a modification of the log-linear models in (1.5) and (1.10) with assumption $P(\epsilon < 0) = F_\epsilon(0) = \alpha$. We can use the scale-mixture model of (1.20) with the modification that $\zeta(u|\theta)$ being the uniform density with support $\{2\theta(\alpha - 1),\ 2\theta\alpha\}$ where θ has an unknown mixing density G. For the sake of brevity, we again omitted the details of the rest of the methodology and related MCMC steps. This model allows only the $(1-\alpha)$-percentile of T to be a log-linear function of covariate.

Our methods can also predict the outcome of a future patient with known covariate values. Our simulation results show that the efficiency gain of semiparametric Bayes estimators is substantial compared to existing frequentist estimators even when our assumptions of (1.5) and (1.10) do not hold (e.g. for Pareto distribution). We do not present any separate simulation study of parametric Bayes estimators because these estimators under diffuse prior information are numerically close to parametric ML estimators. All of these advantages make our proposed methods extremely attractive alternatives to other existing semiparametric method for censored data.

CHAPTER 2

SEMIPARAMETRIC ANALYSIS OF INTERVAL-CENSORED SURVIVAL DATA VIA A LOG-LINEAR MEDIAN REGRESSION MODEL

2.1 Introduction

Analysis of interval censored survival data has become increasingly common problems in many areas including financial, epidemiological, medical, and sociological research studies. A typical example of interval censored data is medical or epidemiological studies of slow-growth diseases that have no immediate outward symptoms (Sun, 2006) [53]. Usually, the occurrence of interval censored observation is mainly due to the nature of the disease and/or the structure of the study design. For such studies (for example, Finkelstein and Wolfe, 1985) [17], the survival time T_i of patient i can not be observed, but can only be determined to be within an interval $(A_i, B_i]$ of a sequence of clinic visit or examination times. Two special cases of interval-censored data are found in current status data (Jewell and van der laan, [28] and others), where either $A_i = 0$ or $B_i = +\infty$. In recent years, there has been a lot of interest and research activity for the models and appropriate analysis for such data. For a more comprehansive review, we refer to the authoritative book by Sun (2006) [53] and the references therein.

Generally, right censored survival data can be seen as a special case of interval censored data, and some of the inference approaches based on right censored data can be applied to interval censored data directly or with some minor modification. However, due to the fundamentally special and complex nature of interval censoring, most of the commonly used survival analysis methods, including methods based on martingale-theory (Anderson et.al., 1992) [1], can not be used for analyzing interval censored survival data. Most of the popular semiparametric models for interval censored survival data focus on modeling the hazard function $h(t \mid x_i)$ given covariate x_i (For example, Sun (2006); Finkelstein and Wolfe (1986); Satten et.al. (1998); Pan (2000); So et.al. (2010)) [53], [18], [48], [43], [52]. For semiparametric Bayesian analysis, there are existing works including Sinha et.al. (1999) [51] and Ghosh & Sinha (2000) [21] dealing with Cox's model [10], and Hanson & Johnson (2004) [22] and Hanson & Yang (2007)[24] using accelerated failure time model. However, for studies with interval-censored data, focusing on the effects of covariate x_i

either on instantaneous risk or change of time-scale may not be appropriate because the design of such studies does not allow continuous monitoring of survival. Also, the Bayesian semiparametric procedures require the knowledge of the prior mean function of the baseline function, either hazard or survival. Often in practice, the available prior information about the study under consideration is only related to certain quantiles, for example the median, of the survival response. This prior information about the median and anticipated range of the change of median for different values of covariate x_i are routinely elicited and then used for power and sample size evaluations of the study (Piantadosi, 2005) [46].

In this chapter, we focus on the inference and theoretical properties of a semiparametric regression survival model with a transform $g_\lambda(\cdot)$ on both the $\log(T_i)$ and regression function $\eta_i = \beta x_i$. We specify a symmetric unimodal nonparametric error distribution for the transformed response $g_\lambda(\log(T))$. This leads to a semiparametric model with log-linear regression function $\exp(\beta x_i)$ for the median of T_i. We develop our inference procedure and associated theoretical justifications within the semi-parametric likelihood as well as full Bayes paradigms. To our knowledge there is no previous works dealing with the median regression function for interval censored survival data. Commonly used self-consistency based and martingale based estimating equations for median regression (Cheng et al., 1997; Yang & Prentice, 1999; McKeague et al., 2001; Bang & Tsiatis, 2002; Portnoy, 2003; Peng & Huang, 2008) [8], [58], [38], [2],[47],[44], have not been extended to deal with interval-censoring. Our computational algorithms for the semiparametric maximum likelihood estimator and the Markov chain Monte Carlo (MCMC) based semiparametric Bayesian estimator are easy to implement. Section 2.2 describes the transformation both-side model and its properties in the context of interval censored data. The semiparametric method is developed in Section 2.3. In Section 2.4, we present an example and make a comparison between different models. We discuss our key findings and comments in Section 2.5.

2.2 Semiparametric Models

For interval censored data, the observations are from subject $i = 1, \ldots, n$ with $\{T_i, x_i\}$, where T_i is random variable representing the failure time of a subject in the study, $x_i = (1, x_{i1}, \ldots, x_{ir})^T$ is the corresponding vector of r covariates along with the intercept term. For interval censored variable T_i, only an interval $(A_i, B_i]$ is observed such that

$$T_i \in (A_i, B_i], \tag{2.1}$$

where $A_i \leq B_i$. In the following chapter, $A_i = B_i$ represents a exact observation and $B_i = \infty$ means a right censored observation. We assume that the interval censoring mechanism is "non-informative", which means the process $\{N_{0_i}(t)\}$ of observation times, clinic visits, and T_i are independent given x_i.

Bickel & Doksum (1981) [3] define a monotone power transformation family, an extension of the Box-Cox power family (Box & Cox, 1964)[4], as

$$g_\lambda(y) = \frac{\text{sign}(y)|y|^\lambda}{\lambda}, \tag{2.2}$$

where $\lambda > 0$ and $\text{sign}(y) = 1$ if $y \geq 0$ and $\text{sign}(y) = -1$ if $y < 0$. Our transform both-side model assumes that $g_\lambda(\log(T_i))$, for an optimal λ, is symmetric unimodal with median

$g_\lambda(\eta_i)$, that is,

$$g_\lambda(\log(T_i)) = g_\lambda(\eta_i) + \varepsilon_i = g_\lambda(\beta x_i) + \varepsilon_i, \tag{2.3}$$

where, ε_i are independent with unimodal and symmetric around zero distribution F_ε. Carroll & Ruppert (1984) and Fitzmaurice et al. (2007) [5] and [19] used a parametric transformation both-side regression model for an uncensored continuous response with the original Box-Cox transformation (Box & Cox, 1964) [4] and parametric normal density for ε_i.

Since $g_\lambda(\cdot)$ is monotonic increasing with inverse $g_\lambda^{-1}(y) = \text{sign}(y)|y\lambda|^{1/\lambda}$, then

$$P[T_i > \exp(\eta_i)] = P[g_\lambda(\log(T_i)) > g_\lambda(\eta_i)] = P[\varepsilon_i > 0] = 1/2 \tag{2.4}$$

as long as ε_i has median 0. As a consequence, the survival time T_i has log-linear median $Q_{0\cdot5} = \exp(\eta_i)$, and any α-percentile of T with $P[T < Q_\alpha] = \alpha$ has the expression

$$Q_\alpha = \exp\{g_\lambda^{-1}(g_\lambda(\eta) + \epsilon_{(\alpha)})\}, \tag{2.5}$$

where $\epsilon_{(\alpha)}$ is the α-percentile of error distribution F_ε with $P[\varepsilon_i < \epsilon_{(\alpha)}] = \alpha$. In some situations, when we are interested in modeling another percentile Q_{α^*} for $\alpha^* \neq 0{\cdot}5$, instead of median $Q_{0\cdot5}$, as a log-linear function $\exp\{\eta(x)\} = \exp(\beta x)$ we can modify (2.3) to assume that unimodal symmetric F_ε satisfies $F_\varepsilon(0) = P[\varepsilon < 0] = \alpha^*$. The expression for other percentiles in (2.5) does not change, except the function $\exp\{\eta(x)\}$ now corresponds to the α^*-percentile of T. For the rest of the article, we deal with only the median functional for the model in (2.3). In the last section, we discuss in more detail the differences and advantages of our model compared to existing survival models, including the quantile regression models of Portnoy (2003), Neocleous et al. (2006) [47], [41] and others.

Theorem 2. *Under non-informative interval-censoring, the parameters (λ, β) of the model (2.3) is identifiable even when F_ε is unknown.*

This important result about identifiability is true even for the restricted case of current status data, when either $A_i = 0$ or $B_i = +\infty$. The proof is based on the fact that $S(t \mid x = 0)$ and $S(t \mid x = 1)$ are identifiable from current status data. Medians of $S(t \mid x = 0)$ and $S(t \mid x = 1)$ identify the parameters (λ, β). We would also like to mention that, for model (2.3), there exists a unique $(\lambda, \beta, F_\varepsilon)$ satisfying (2.3). The proof is omitted.

Also, the sign and magnitude of any component of β determines the relationship between every percentile of T and that component of covariate.

Proposition 1. $Q_{\alpha_1}(x) \geq Q_{\alpha_1}(x^*) \Leftrightarrow Q_{\alpha_2}(x) \geq Q_{\alpha_2}(x^*)$, *for all* (α_1, α_2).

The ordering of the percentiles for any two subjects remain same for all α. Without loss of generality, to show this, we take a covariate vector $x = (x_1, x_2)$ with two scalar components. For (2.3), $\beta_1 > 0 \Rightarrow Q_\alpha(x_1, x_2) > Q_\alpha(x_1^*, x_2)$ for all α as long as $x_1 > x_1^*$ where $Q_\alpha(x_1, x_2)$ is the α-percentile of the survival time for subject with covariate value $x = (x_1, x_2)$. This model property is similar to the uniform ordering of the survival functions property of the Cox model.

2.3 Model Estimation and Inference

For the model in (2.3), the likelihood contribution of each subject is $P(T_i \in (A_i, B_i]) = P(\varepsilon \in (\widetilde{A}_i, \widetilde{B}_i])$, where $\widetilde{A}_i = \mathrm{g}_\lambda(\log(A_i)) - \mathrm{g}_\lambda(\eta_i)$ and $\widetilde{B}_i = \mathrm{g}_\lambda(\log(B_i)) - \mathrm{g}_\lambda(\eta_i)$. The likelihood based on the observed interval censored data $\mathcal{D}$ is now

$$L(\lambda, \beta, F_\varepsilon \mid \mathcal{D}) = \prod_{i=1}^{n} P(\varepsilon_i \in (\widetilde{A}_i, \widetilde{B}_i]) = \prod_{i=1}^{n} \left[F_\varepsilon(\widetilde{B}_i) - F_\varepsilon(\widetilde{A}_i) \right], \tag{2.6}$$

where $\mathcal{D} = \{T_i \in (A_i, B_i], x_i : i = 1, \ldots, n\}$. An appropriate parametric distribution F_ε, for example, a normal with mean 0 and variance σ^2, results in a parametric likelihood for (2.6). A numerical method, such as Nelder & Mead (1965) [40] can be used to obtain the maximum likelihood estimator $(\widehat{\lambda}, \widehat{\beta}, \widehat{\sigma})$ of (λ, β, σ), and the corresponding consistent variance of estimate as the inverse of the observed Fisher information matrix. For the parametric Bayesian analysis, the posterior density is $p(\lambda, \beta, \sigma \mid \mathcal{D}) \propto L(\lambda, \beta, \sigma \mid \mathcal{D})\pi(\lambda, \beta, \sigma)$, where $\pi(\lambda, \beta, \sigma)$ is the joint prior density.

The parametric assumption about F_ε may be inappropriate and restrictive in practice. The class of all unknown unimodal symmetric distribution F_ε can be expressed as a scale-mixture of uniforms

$$F_\varepsilon(\epsilon) = \int_0^\infty F_U(\epsilon \mid \theta) \mathrm{d}G(\theta), \tag{2.7}$$

for some mixing distribution $G(\theta)$, where $F_U(u \mid \theta)$ is a uniform distribution with support $(-\theta, +\theta)$ for $\theta > 0$ (Feller, 1971) [14]. The full semiparametric likelihood of (λ, β, G) can be derived as

$$L(\lambda, \beta, G \mid \mathcal{D}) \propto \prod_{i=1}^{n} \left[\int_0^{+\infty} \int_{\widetilde{A}_i}^{\widetilde{B}_i} f_U(\epsilon \mid \theta) \ \mathrm{d}\epsilon \ \mathrm{d}G(\theta) \right]$$

from (2.6) and (2.7). For the semiparametric-likelihood analysis, we use an "empirical" version of the above likelihood, where $F_\varepsilon(\epsilon)$ in (2.7) is replaced with

$$F_\varepsilon(\epsilon) = \sum_{j=1}^{K} p_j F_U(\epsilon \mid \theta_j), \tag{2.8}$$

where the mixing distribution $G(\theta)$ is discrete with finite support $\Theta = \{\theta_1, \ldots, \theta_K\}$, with unknown $0 < \theta_1 < \cdots < \theta_K$. Maximizing this likelihood of $(\lambda, \beta, \Theta, p)$ is tantamount to maximizing the following likelihood with $K \leq n$.

Theorem 3. *For discrete $G(\cdot)$ with support Θ and $\mathrm{pr}(\theta = \theta_j) = p_j$ for $0 \leq p_j \leq 1$ and $\sum_{j=1}^{K} p_j = 1$, the log-likelihood of (2.6) is*

$$l(\lambda, \beta, F_\varepsilon \mid \mathcal{D}) = l(\lambda, \beta, p, \Theta) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{K} p_j \frac{\widetilde{B}_{ij} - \widetilde{A}_{ij}}{2\theta_j} \right] \tag{2.9}$$

where $\widetilde{A}_{ij} = \min\{\max\{-\theta_j, \widetilde{A}_i\}, \theta_j\}$, $\widetilde{B}_{ij} = \max\{\min\{\theta_j, \widetilde{B}_i\}, -\theta_j\}$.

See Appendix B for the proof. Using results of Wong & Severini (1991) [57], under similar regularity conditions, the maximum likelihood estimator $\widehat{\beta}$ of the regression parameter has $n^{1/2}$ convergence rate and is asymptotically efficient. This is particularly true because $F_\varepsilon(\epsilon)$ defined in (2.8) has a density function. The maxima of the likelihood in (2.6) always exists because this likelihood function is a product of probabilities, a bounded function with $0 \leq L(\lambda, \beta, F_\varepsilon \mid \mathcal{D}) \leq 1$.

Computing the maximum likelihood estimator of $(\lambda, \beta, \Theta, p)$ via directly maximizing (2.9) may be computationally intensive. To reduce the computational burden, we propose the use of following iterative procedure with two steps in each iteration. At each iteration, we begin with the most recent value of $(\widehat{\lambda}, \widehat{\beta})$ and $(\widehat{\Theta}, \widehat{p})$. We suggest using the parametric maximum likelihood estimator as the initial value of $(\widehat{\lambda}, \widehat{\beta})$ for the first iteration.

Step 1: Maximize (2.9) with respect to (Θ, p) to obtain the current value of $(\widehat{\Theta}, \widehat{p})$, where $(\lambda, \beta) = (\widehat{\lambda}, \widehat{\beta})$ is fixed.

At this step, the optimal Θ is the set of ordered distinct values of $\{|\widetilde{A}_i|, |\widetilde{B}_i| : i = 1, \ldots, n\}$, where $\widetilde{A}_i$ and $\widetilde{B}_i$ are known functions of $(\widehat{\lambda}, \widehat{\beta})$, x_i and (A_i, B_i). This implies that the only unknown parameter at this stage is the vector p. There is unique maxima of $l(\widehat{\lambda}, \widehat{\beta}, p, \Theta \mid \mathcal{D})$ because it is a concave function of p.

Step 2: Considering the values of $(\widehat{\Theta}, \widehat{p})$ obtained in Step 1 as fixed, maximize the likelihood in (2.9) as a function of (λ, β) to obtain a new $(\widehat{\lambda}, \widehat{\beta})$. Go back to step 1 and continue the iterations until convergence.

This can be implemented using nonlinear optimization algorithms such as Nelder & Mead (1965) [40]. This iterative procedure, under mild conditions, maximizes the profile likelihood (Murphy & van der Vaart, 2000)[39]

$$l_P(\lambda, \beta \mid \mathcal{D}) = \max_{(p,\Theta)} l(\lambda, \beta, p, \Theta \mid \mathcal{D}),$$

even though $l_P(\lambda, \beta \mid \mathcal{D})$ does not have any closed form expression in this case. The (j, k) component of the estimated limiting covariance matrix of $\widehat{\beta}$ is given as

$$-\mathcal{E}_n^2 \Big[l_P(\widehat{\lambda}, \widehat{\beta} + \mathcal{E}_n u_j + \mathcal{E}_n u_k) - l_P(\widehat{\lambda}, \widehat{\beta} + \mathcal{E}_n u_j - \mathcal{E}_n u_k) \\ - l_P(\widehat{\lambda}, \widehat{\beta} - \mathcal{E}_n u_j + \mathcal{E}_n u_k) + l_P(\widehat{\lambda}, \widehat{\beta}) \Big],$$

where u_j is the standard basis vector with 1 in component j, and $\mathcal{E}_n$ is a scalar of order $n^{-1/2}$ (Murphy & van der Vaart, 2000) [39].

Standard errors for the maximum likelihood estimate $\hat{\beta}$ are obtained using the inverse of the observed information matrix obtained via numerical differentiation of the likelihood in (2.9). When the sample size is large, the implementation of step-1 of the algorithm may turn out to be computationally difficult. In this case, we recommend using a penalty function of smoothness on p.

Another option for semiparametric analysis is to use a full Bayes procedure based on Markov chain Monte Carlo samples from the joint posterior

$$\text{pr}(\lambda, \beta, F_\varepsilon \mid \mathcal{D}) \propto L(\lambda, \beta, F_\varepsilon \mid \mathcal{D}) \pi_1(\lambda, \beta) \pi_2(F_\varepsilon),$$

where $\pi_1(\lambda, \beta)$ and $\pi_2(F_\varepsilon)$ are the independent, a reasonable assumption of convenience, priors for the parametric part (λ, β) and the nonparametric F_ε, respectively. This full Bayes semiparametric model uses the scale-mixture of uniforms in (2.7) as the class of symmetric unimodal F_ε, with no requirement on the unknown mixing distribution G being discrete. We use a Dirichlet process (Ferguson, 1973) [15] prior $G \sim \mathrm{DP}(G_0, \alpha)$ for unknown $G(\theta)$, where $G_0(\cdot)$ is the known prior mean of $G(\cdot)$ and $\alpha > 0$ is the precision around mean G_0. The following theorem gives us a method of choosing $G_0(\cdot)$ based on the desired form of prior mean $F_0(\cdot)$ of $F_\varepsilon(\cdot)$. Typically, $F_0(\cdot)$ is assumed to be a known, specified a priori, parametric distribution function with corresponding density $f_0(\cdot)$.

Theorem 4. *When* $\mathrm{E}_{prior}[F_\varepsilon(\cdot)] = F_0(\cdot)$*, for* $F_0(\cdot)$ *symmetric around zero with corresponding density* $f_0(\cdot)$*, the corresponding prior mean* $G_0(u)$ *of* $G(\cdot)$ *in (2.7) is* $\mathrm{d}G_0(u) = -2u\mathrm{d}f_0(u)$*,* $u > 0$*.*

The proof follows from a result by Khintchine (1938) [30]. To ensure the prior mean of F_ε as $N(0, v^2)$, the functional form of $G_0(u)$ has to be Gamma$(3/2, 1/(2v^2))$. For the implementation of the Markov chain Monte Carlo tool, we use a finite approximation of the constructive definition of the Dirichlet process mixture prior process (Sethuraman, 1994) [49] as $F_\varepsilon(\epsilon) \cong \sum_{j=1}^{K} F_U(\epsilon \mid \theta) p_j$, where $\theta_j \sim G_0(\cdot)$, $p_j = V_j \prod_{\ell<j}(1 - V_\ell)$, $V_j \sim \mathrm{Beta}(1, \alpha)$. We use the WinBUGS software (Lunn et al., 2000) [37] to obtain the Markov chain Monte Carlo samples of the posterior distribution $p(\lambda, \beta, F_\varepsilon \mid \mathcal{D})$, and implement a semi-parametric Bayesian analysis in Section 2.4.

2.4 Data Example

We illustrate our methods via reanalysis of a retrospective study on early stage breast cancer patients (Table 2.1), who had been treated at the Joint Center for Radiation Therapy in Boston between 1976 and 1980. The data are reported in Finkelstein & Wolfe (1985) [17] with two treatment arms: $x_{1i} = 0$ if patient i received RT (radiation therapy) and $x_{1i} = 1$ if she received RT+CH (radiation therapy + adjuvant chemotherapy), where $x_i = (1, x_{1i})$ and regression parameter $\beta = (\beta_0, \beta_1)$, so that the median regression model is $Q_{0.5} = \exp(\beta_0 + \beta_1 x_{1i})$. The patients were observed in irregular intervals, mostly of 4 to 6 months, for the event of cosmesis. In this study, 46 patients received radiation therapy (RT) and 48 patients received radiation therapy plus adjuvant chemotherapy (RH+CH). The goal of this study is to compare the effect of two treatment arms, RH and RH+CH.

The maximum likelihood estimate of the parameters $(\lambda, \beta_0, \beta_1)$ for the parametric model with Gaussian F_ε and for the semiparametric model with discrete scale-mixture of uniforms for F_ε are given in Table 2.2. Figure 2.1 presents the estimated survival functions of two treatment groups under the parametric model. The horizontal lines are nonparametric estimators (Peto, 1973) [45] of survival functions, estimated separately for two treatment arms. Based on the maximum likelihood estimator of the semiparametric model, Figure 2.2 presents the estimated error density of (2.3) and the corresponding estimated survival functions of the two groups. This figure also presents the nonparametric estimators for the two treatment arms, for comparison with semi-parametric estimators. We observe that the semiparametric estimators of the survival curves matches the nonparametric estimators

Table 2.1: Observed intervals in months for times to breast retraction of early breast cancer patients (Sun, 2006).

Observed intervals in months
RT (radiation therapy):
(45,], (25,37], (37,], (4,11], (17,25], (6,10], (46,], (0,5], (33,], (15,],
(0,7], (26,40], (18,], (46,], (19,26], (46,], (46,], (24,], (11,15], (11,18]
(46,], (27,34], (36,], (37,], (22,], (7,16], (36,44], (5,12], (38,], (34,]
(17,], (46,], (19,35], (46,], (5,12], (9,14], (36,48], (17,25], (36,], (46,]
(37,44], (37,], (24,], (0,8], (40,], (33,]
RCT (radiation therapy + adjuvant chemotherapy):
(8,12], (0,5], (30,34], (16,20], (13,], (0,22], (5,8], (13,], (30,36], (18,25]
(24,31], (12,20], (10,17], (17,24], (18,24], (17,27], (11,], (8,21], (17,26], (35,]
(17,23], (33,40], (4,9], (16,60], (33,], (24,30], (31,], (11,], (15,22], (35,39]
(16,24], (13,39], (15,19], (23,], (11,17], (13,], (19,32], (4,8], (22,], (44,48]
(11,13], (34,], (34,], (22,32], (11,20], (14,17], (10,35], (48,]

better than the match of the parametric estimators in Figure 2.1. Using the parametric maximum likelihood estimator and the semiparametric maximum likelihood estimator, Table 2.3 presents the point and 95% interval estimates of the medians of the RT ($\exp(\beta_0)$) and RT+CH ($\exp(\beta_0+\beta_1)$) treatment arms, as well as the ratio of these two medians ($\exp(-\beta_1)$). For both parametric and semiparametric likelihood methods, the interval estimates of the medians and their ratios use the estimated standard errors obtained via the delta-method based on the estimated variance-covariance matrix of the regression parameter estimate $\widehat{\beta}$.

For the Bayesian analysis, we assume the parameters to be a priori independent with joint prior

$$\pi(\lambda, \beta_0, \beta_1, F_\varepsilon) \propto \pi_1(\lambda)\pi_2(\beta_0)\pi_3(\beta_1)\pi_4(F_\varepsilon).$$

It is possible to elicit informative priors for β_0 and β_1 using the prior information, when

Table 2.2: Maximum Likelihood estimative of regression parameters β for transformation both-side model.

Parametric	Estimate	SE	95% interval estimate
λ	1·629	0·677	(0·302, 2·956)
β_0	3·551	0·133	(3·289, 3·813)
β_1	-0·399	0·183	(-0·759, -0·039)
Semiparametric	Estimate	SE	95% interval
λ	0·968	0·031	(0·908, 1·028)
β_0	3·666	0·072	(3·517, 3·808)
β_1	-0·544	0·101	(-0·743, -0·345)

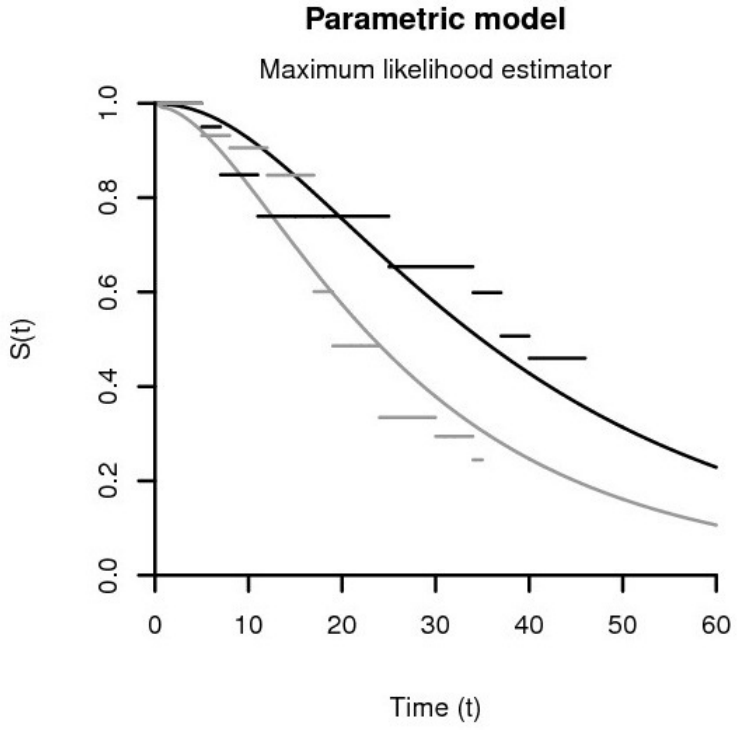

Figure 2.1: Parametric maximum likelihood estimator survival functions for two treatment arms, black for RT and gray for RT+CH, horizontal lines are Peto's nonparametric estimators.

Table 2.3: Estimated medians using maximum likelihood estimator.

Parametric	Point estimate	95% interval estimate
RT Group	34·87	(26·84, 45·32)
RT+CH Group	23·39	(18·05, 30·30)
Ratio of medians	1·49	(1·04, 2·13)
Semi-parametric	Point estimate	95% interval estimate
RT Group	39·10	(33·68, 45·06)
RT+CH Group	22·69	(16·12, 31·93)
Ratio of medians	1·72	(1·41, 2·10)

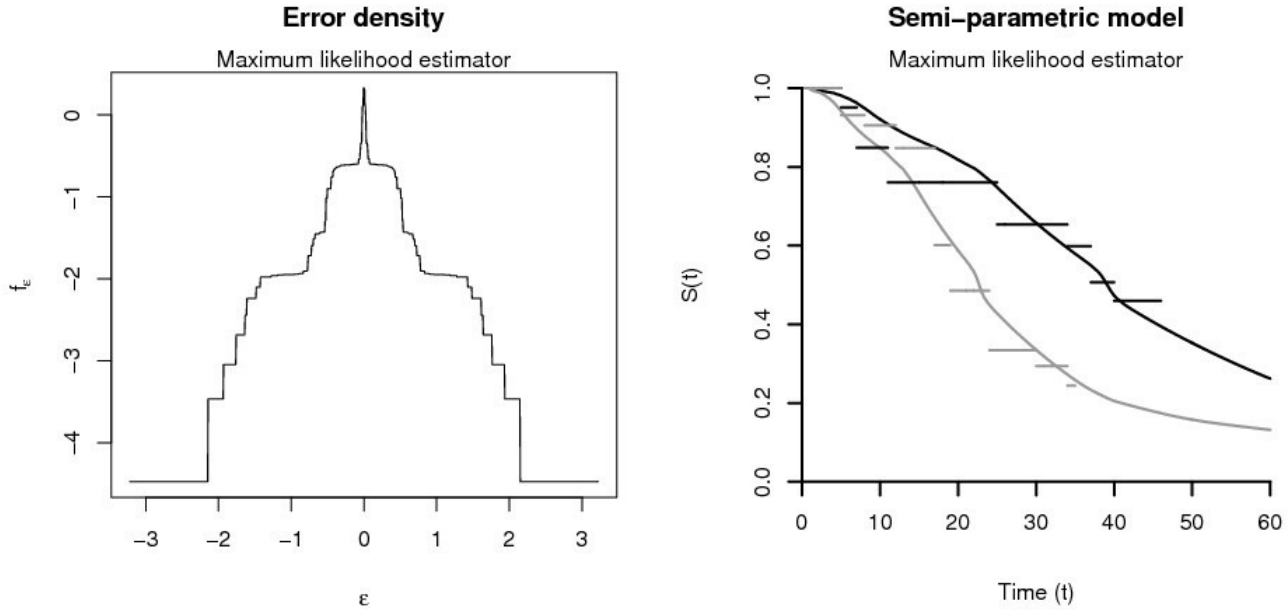

Figure 2.2: Semi-parametric maximum likelihood estimator of: (a) error density and (b) survival functions for two treatment arms, black for RT and gray for RT+CH, horizontal lines are Peto's nonparametric estimators.

Table 2.4: Bayesian estimative of transformation both-side model.

Parametric	Mean	SE	95% credible interval
λ	1·260	0·300	(0·692, 1·808)
β_0	3·547	0·149	(3·266, 3·821)
β_1	-0·406	0·204	(-0·760, 0·039)
Semi-parametric	Mean	SE	95% credible interval
λ	1·054	0·132	(0·807, 1·315)
β_0	3·661	0·158	(3·380, 3·974)
β_1	-0·501	0·202	(-0·920, -0·130)

they are available about the possible support and the prior guess of the median survival times $\exp(\beta_0+\beta_1)$ and $\exp(\beta_0)$ of two treatment arms. Reviewing the literature, we found no such prior information about these median survival times. We instead used fairly non-informative and skeptical priors, a $N(0,5^2)$ for β_0 , and a $N(0,1^2)$ for β_1. It is reasonable to assume a priori that the transformation parameter λ has an effective range in the interval $(0,4)$. Box & Cox (1964) [4] themselves cautioned against using $\lambda >> 3$ due to lack of any reasonable physical interpretation of the model. We suggest using a prior density with mean 1, because $\lambda = 1$ means that no transformation is needed for achieving the symmetry for the log-survival time. In this paper, we are using the gamma prior with mean 1 and variance 1/2. For our analysis, we use the same priors of the parameters (β,λ) for the parametric and semiparametric Bayes analysis.

For the parametric Bayes analysis, we need to specify a prior for the error-variance σ to assign our prior opinion π_4 for Gaussian F_ε. We use a gamma prior with mean 1 and variance 1/2 for σ. For semiparametric Bayes, we use a Gamma distribution with mean 4 and variance 8 as G_0, the prior mean of G. Using Theorem 4, this corresponds to the double-exponential density with scale 0·5 as the prior mean F_0 of F_ε. Also, we use $\alpha = 1$, which means that the precision of the prior opinion about unknown G is very low.

The Bayesian estimates, posterior mean and 95% credible interval, of the parameters (β_1,λ) for parametric and semiparametric models are given in Table 2.4. Figure 2.3 presents the posterior means of the parametric survival functions of two treatment arms. Figure 2.4 presents the posterior means of the error density F_ε and of the two survival functions under the semiparametric model. Table 2.5 presents the Bayesian estimates of the median survival times of two groups and the ratio of two medians under these two models. We evaluate the posterior probability of $\beta_1 < 0$ given observed data. For the parametric model, $P(\beta_1 < 0 \mid \mathcal{D}) \cong 0{\cdot}98$, and for the semiparametric model, $P(\beta_1 < 0 \mid \mathcal{D}) \cong 0{\cdot}978$. The convergence diagnostics of the Markov chain Monte Carlo samples were monitored using trace plots and plots of others standard diagnostics.

The results obtained under model (2.3) can be compared to those obtained under Cox's model $S_1(t \mid x) = \{S_0(t)\}^{\exp(\eta x)}$ only when $S_0(t)$ is exponential. In this special case of Cox's model, the median survival time for covariate x is $Q(x) = \log(2)/\{\nu\exp(\eta x)\}$, where $S_0(t) = \exp(-\nu t)$, and the ratio of medians of two treatment arms, $x = 1$ versus $x = 0$, is equal to $\exp(-\eta)$. The corresponding estimates of $\exp(-\eta)$ obtained by Sinha et al.

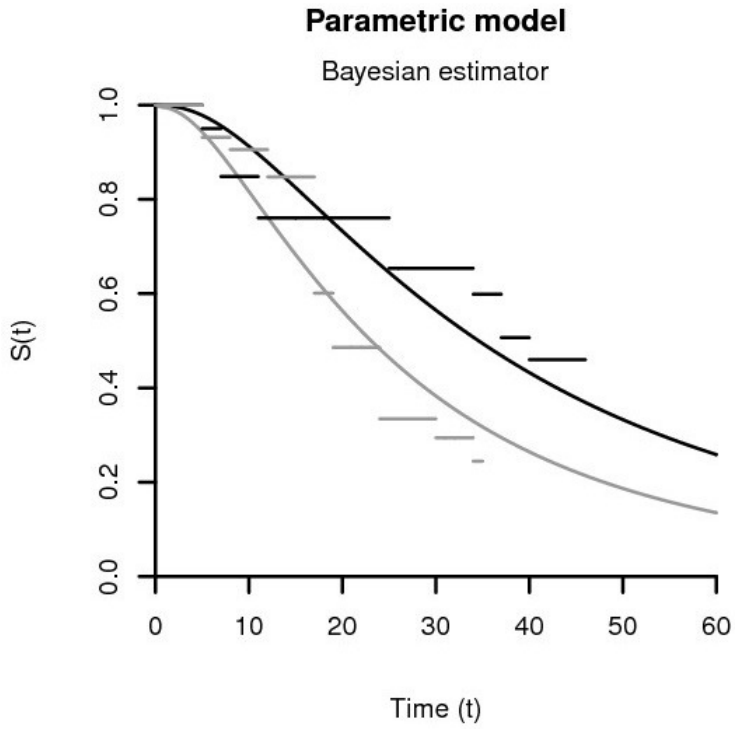

Figure 2.3: Parametric posterior mean of survival functions for two treatment arms, black for RT and gray for RT+CH, horizontal lines are Peto's nonparametric estimators.

Table 2.5: Estimated medians using BE.

parametric	Point estimate	95% credible interval
RT Group	35·12	(28·20, 45·64)
RT+CH Group	23·34	(17·84, 30·16)
Ratio of medians	1·53	(0·96, 2·13)
Semi-parametric	Point estimate	95% credible interval
RT Group	39·40	(28·85, 52·40)
RT+CH Group	23·73	(18·37, 28·77)
Ratio of medians	1·68	(0·89, 2·22)

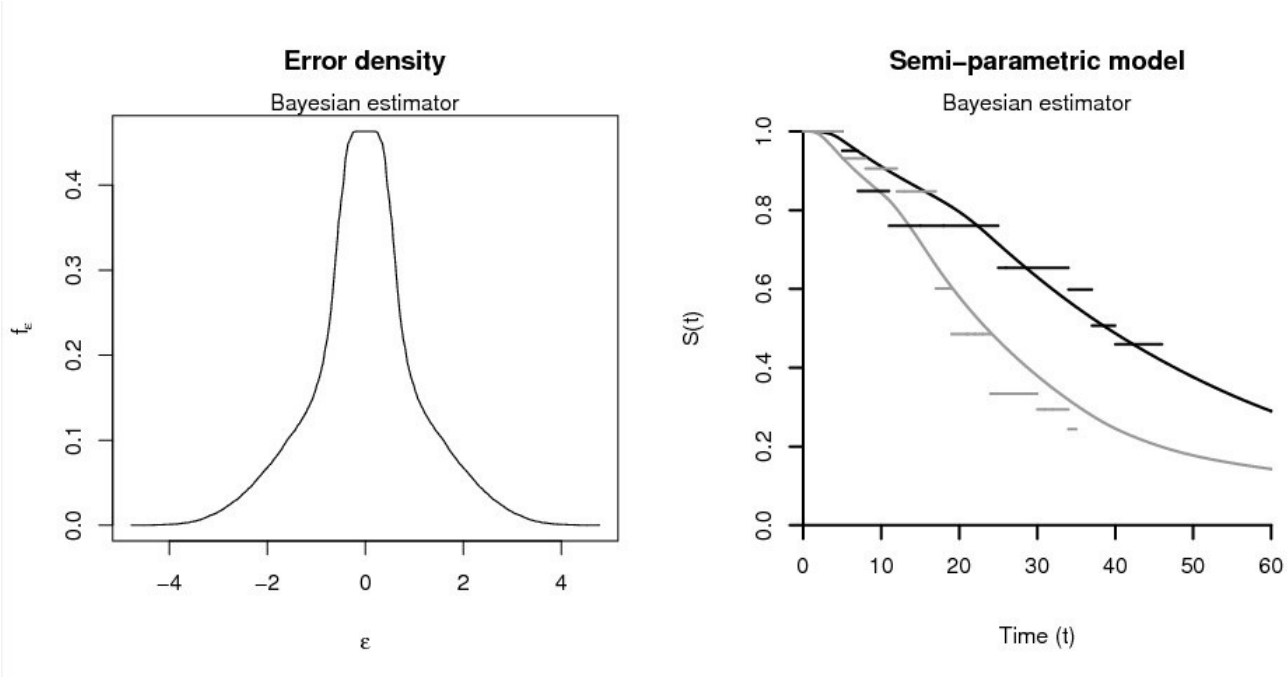

Figure 2.4: Semiparametric posterior mean of (a) error density and (b) survival functions for two treatment arms, black for RT and gray for RT+CH, horizontal lines are Peto's nonparametric estimators.

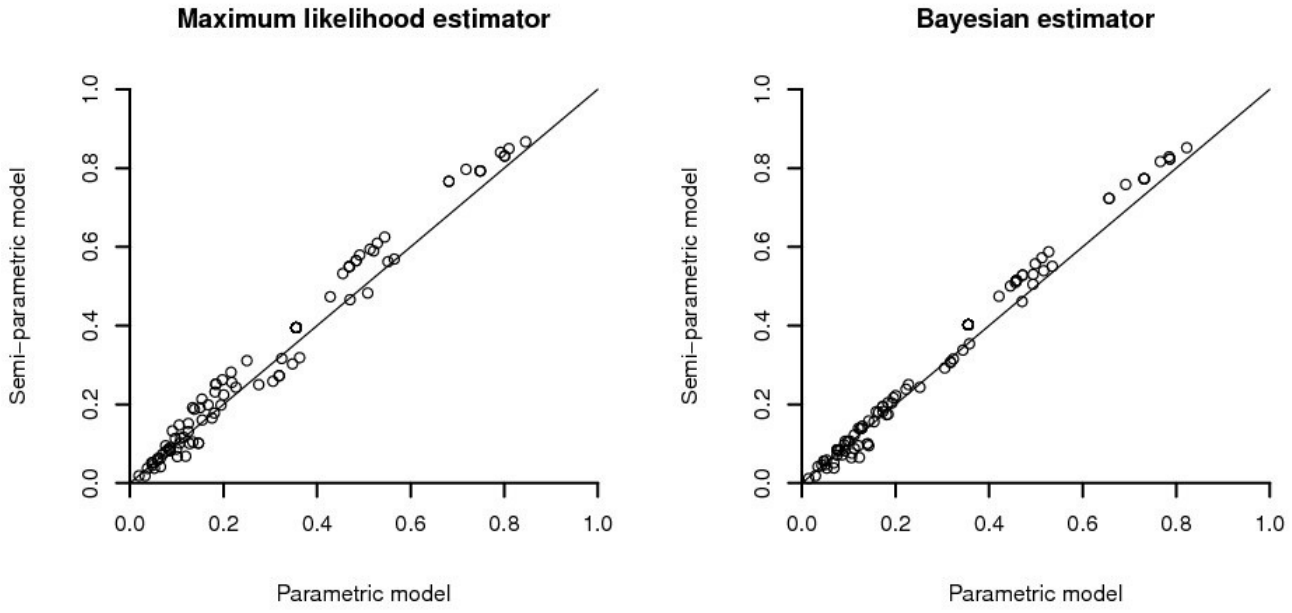

Figure 2.5: Comparison between parametric and semi-parametric models: (a) maximum likelihood estimator and (b) Bayes estimator.

(1999)[51] and by Finkelstein & Wolfe (1985) [17], are similar, in values, to our maximum likelihood and Bayes estimators of $\exp(\beta_1)$. Overall, there is high posterior evidence to conclude that the median time to observe the cosmetic effects was lower in the patients under the RT+CH treatment than the corresponding median for RT alone.

For the likelihood analyses, we compare the parametric and semiparametric model estimates of the probability $q_i = P(T \in (A_i, B_i] \mid x_i)$ of the observed data from subject i. In Figure 2.5 we present a plot to compare the maximum likelihood estimators for the parametric and semiparametric models. For Bayesian analysis, we also present a similar plot, however, it is based on cross-validated posterior probability $\mathrm{E}[q_i \mid D_{-i}]$ (Gelfand et al., 1992)[20], where D_{-i} is the data based on observed data minus the observation from patient i. For both plots, we see that most of the points, around 70%, are above the 45^o line, implying that the semiparametric model fits the data better under both methods. In both figures 2.2 and 2.4, the semiparametric estimators of the survival curves under model (2.3) show good fidelity to the nonparametric estimators. This, supports a better fit of the model (2.3) for this data compared to the apparent lack of fit of Cox's model, as mentioned in Sinha et al. (1999) [51] among others.

2.5 Final remarks

The transformation both-side model has some advantages over the other existing methods of inference for interval censored data. The model can focus on the median and quan-

tiles which are more appropriate for continuously monitored studies, than instantaneous risk. The semiparamteric estimation give us a smooth continuous estimated survival functions; The median and any other quantile of the survival time can be obtained from the estimated parameters of the model. In existing quantile regression models (For example, Portnoy (2003) [47]), every quantile is assumed to be a linear function $Q_\alpha(x) = \beta_\alpha x$ for every $0 < \alpha < 1$, where $P[T < Q_\alpha(x)] = \alpha$. When x is unbounded, this implies that these linear functions are parallel to each other. In model (2.3), only one quantile of interest, say the median, is a linear function. The computation of the maximum likelihood estimator involves an iterative algorithm with two simple finite-dimensional maximization steps within each iteration; for semiparametric Bayesian analysis, we only use a Markov chain Monte Carlo technique, implementable via WinBUGS. The hazard functions for this model can be non-monotone. Model (2.3) can be also written as a location family $Y = \log(T) = \eta + \epsilon$, where $\eta = \beta x$. However, unlike the accelerated lifetime model, the distribution function $F_\varepsilon(\mathrm{g}_\lambda(\epsilon + \eta) - \mathrm{g}_\lambda(\eta))$ of the error ϵ depends on covariate x, a heteroscedastic location family model. The data analysis illustrate the performance of the model, computational and interpretational conveniences as well as the ease of model diagnostics.

For large datasets, the value of K, the number of uniforms used in (2.8), can be large for the likelihood. For a comparison, in our experience, we found that 7 to 8 components is large enough to achieve a good approximation for the Sethuraman's construction used in Bayes computation. Thus, using 7 to 8 components, we have found our approach to be computationally feasible for a large variety of datasets.

CHAPTER 3

REGULARIZED MEDIAN REGRESSION AND OUTLIER DETECTION VIA TRANSFORM-BOTH-SIDES MODEL

3.1 Introduction

We consider the following general linear regression model of the form:

$$\mathbf{Y} = \mu \mathbf{1}_n + \mathbf{X}\beta + \epsilon \ , \tag{3.1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the $n \times 1$ vector of response, μ is the overall mean, $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ is the $n \times p$ matrix of covariates and ϵ is the $n \times 1$ vector of iid normal errors with mean 0 and variance σ^2.

The Least Absolute Shrinkage and Selection Operater (LASSO) is proposed by Tibshirani (1996) ([54])for simultaneous variable selection and parameter estimation. This procedure is constructed within the penalized likelihood framework and shrinks some regression coefficients towards exactly zero. The lasso minimizes the residual sum of square with a constraint which is expressed in term of the L_1-norm of the coefficient vector of β:

$$\hat{\beta}_L = \arg\min_{\beta} (\tilde{\mathbf{y}} - \mathbf{X}\beta)'(\tilde{\mathbf{y}} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \ , \tag{3.2}$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}_n$ is the centered observed response vector and $\lambda \geq 0$ is the tuning parameter of penalty. When $\lambda = 0$, $\hat{\beta}_L$ is the ordinary least-square estimate, and it would be shrinked towards zero when λ is sufficiently large. The lease angle regression algorithm (LARS) (Efron et al., 2004; Osborne et al.,2000) ([11])([42]) provides an efficient implementation of LASSO computation.

However, if the normality assumption of (3.1) does not hold, penalized median/quantile regression provides a useful alternative to classical LASSO estimates for its superior robustness properties, richer information and better prediction accuracy. The natural extension of LASSO penalty in quantile regression context can be defined as (Koenker, R. 2005)([31]):

$$\hat{\beta}_L(\tau) = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(\tilde{y}_i - X_i\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \ , \tag{3.3}$$

where $\rho_\tau(\cdot)$ is the loss function

$$\rho_\tau(t) = \begin{cases} \tau t, & \text{if } t > 0 \\ -(1-\tau)t, & \text{if } t < 0 \end{cases}$$

The above penalized estimators (3.3) of quantile LASSO can be easily solved by standard linear programming techniques.

Most of the quantile LASSO based on loss function is very hard to accommodate heteroscedasticity and outlier detection, which are common phenomenas in statistical data analysis. Often in real data analysis, covariate X may affect the scale and shape of the distribution of random noise variable ϵ. Thus, the possible existence of heteroscedasticity can cause big concern in the regression analysis and model selection. A model which accommodates heteroscedasticity can provide a very useful tool to this problem. Another pervasive problem in analyzing the routine data set is the outliers detection. Surprisingly, outlier often go unnoticed in practice, while undetected outlier would possibly result in serious bias in parameter estimation and model selection. Our Goal in this work is to develop a model which is able to account for the heteroscedasticity of $\mathbf{Y}$, together with robust parameter estimation and outlier detection using sparsity penalization.

The rest of paper proceeds as follows. In Section 2, we introduce the proposed regularized median regression via transform-both-sides Model. An iterative procedure is developed to get estimators. A extended model to accommodate gross outliers is also discussed here. In section 3, we compare the proposed median lasso to the frequentist counterparts via simulation studies. In section 4, we have the final discussion and conclusion.

3.2 Regularized Median Regression

Bickel and Doksum (1981)([3]) proposed a monotone power transformation, an extension of the Box-Cox power family,

$$b_\lambda(y) = \frac{y^\lambda \text{sign}(y) - 1}{\lambda}, \text{ for } \lambda > 0 \ , \tag{3.4}$$

where $\text{sign}(y) = 1$ if $y \geq 0$ and $\text{sign}(y) = -1$ if $y < 0$. We assume that under an optimal λ, the transformed response $b_\lambda(y_i)$ has symmetric and unimodal distribution with mean $b_\lambda(\sum_j x_{ij}\beta_j)$ and variance σ^2, that is

$$b_\lambda(y_i) = b_\lambda(\sum_j x_{ij}\beta_j) + \epsilon_i \ , \tag{3.5}$$

where ϵ_i are i.i.d. $\sim F$. The transformation $b_\lambda(y)$ in (3.5) is monotone with derivative $b'_\lambda(y) = |y|^{\lambda-1}$. The median of y_i is $M_i = \sum_j x_{ij}\beta_j$ because $P[y_i > M_i] = P\left[b_\lambda\{y_i\} > b_\lambda(M_i)\right] = F_\epsilon(0) = 1/2$, where F_ϵ is the cdf of ϵ. As a consequence, the observation y_i has a linear median regression function $Q_{0.5}(x_{ij}) = M_i = x_{ij}\beta_j$.

Usually we introduce an additional scale parameter σ. For example, if the transformed model is Gaussian, we have ϵ_i are i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. When $F \sim \mathcal{N}(0, \sigma^2)$. Suppose for a

vector $\boldsymbol{a}$, $b_\lambda(\boldsymbol{a})$ is applied componentwise. By assume the sparsity on $\boldsymbol{\beta}$, we can use an L_1 -penalized regression to minimize

$$\frac{1}{2}\|b_\lambda(\boldsymbol{y}) - b_\lambda(\boldsymbol{X}\boldsymbol{\beta})\|_2^2 + \sum_{j=1}^{p} P(\beta_i;\mu) =: F(\boldsymbol{\beta}) \tag{3.6}$$

by an thresholding-(denoted by Θ) based iterative procedure (She, 2011) ([50]). A modified version is described in Algorithm 1. More details on the theorems are discussed in the Appendix C. (Thanks to Dr. She for his tremendous contribution and help of the theorems)

Algorithm 1 Modified Θ-IPOD using an L_1 penalty on $\boldsymbol{\beta}$

Require: $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \mathbb{R}^n$, $\mu > 0$, $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$, and threshold $\Theta(.;.)$
Ensure: $j \leftarrow 0, \boldsymbol{\beta}^{(j)} \leftarrow \boldsymbol{\beta}^{(0)}$, converged $\leftarrow$ FALSE
  while not converged **do**
    $\boldsymbol{\beta}^{(j+1)} \leftarrow \Theta\left(\boldsymbol{\beta}^{(j)} + \boldsymbol{X}^T \mathrm{diag}\{b'_\lambda(\boldsymbol{X}\boldsymbol{\beta}^{(j)}\}[b_\lambda(\boldsymbol{X}\boldsymbol{\beta}^{(j)}) - b_\lambda(\boldsymbol{y})]; \mu\right)$
    converged $\leftarrow \| \boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)} \|_\infty < \varepsilon$
    $j \leftarrow j + 1$
  end while
  Return $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(j)}$.

The convergence is guaranteed *globally*, i.e., the objective function $F(\cdot)$ is decreasing during the iteration given any initial point, as long as the 2-norm $\|\boldsymbol{X}\|_2$ is small enough. This can be achieved by performing a scaling $\boldsymbol{X} \leftarrow \boldsymbol{X}/K$ with a large value of K. We can theoretically evaluate K under some regularization conditions. Specifically, let $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta}$, and assume all components of $\boldsymbol{y}$ are bounded away from zero (say by δ) as well as the components of $\boldsymbol{z}(\boldsymbol{\beta})$ for any iterate. The Hessian of the first term in F is then

$$\boldsymbol{X}^T \mathrm{diag}\left\{b'^2_\lambda(\boldsymbol{z}) + (b_\lambda(\boldsymbol{z}) - b_\lambda(\boldsymbol{y}))b''_\lambda(\boldsymbol{z})\right\}\boldsymbol{X},$$

where, again, $b'^2_\lambda, b''_\lambda, b_\lambda$ are applied elementwise. Then the bound of the diagonal entries for the given data gives K^2.

The Θ is a threshold function satisfying

$$P(t;\mu) - P(0;\mu) = \int_0^{|t|} (\sup\{s : \Theta(s;\mu) \le u\} - u)\,\mathrm{d}u + q(t;\mu) \tag{3.7}$$

for some nonnegative $q(\cdot;\mu)$ satisfying $q(\Theta(t;\mu);\mu) = 0$ for any $t \in \mathbb{R}$. independent of μ.

In this article, we use the following soft thresholding:

$$\Theta_{soft}(t;\mu) = \begin{cases} 0, & \text{if } |x| \le \lambda \\ x - \mathrm{sgn}(x)\lambda, & \text{if } |x| > \lambda. \end{cases} \tag{3.8}$$

The iterative Algorithm 1 can deal with the l_1 penalty, l_p-penalties, and the $l_0 + l_2$

penalty

$$P_1(t;\mu) = \mu|t|, \tag{3.9}$$

$$P_{l_p}(t;\mu) = \mu|t|^p, \quad 0 < p < 1 \tag{3.10}$$

$$P(t;\mu,\eta) = \frac{1}{2}\frac{\mu^2}{1+\eta}1_{x\neq 0} + \frac{1}{2}\eta x^2. \tag{3.11}$$

The last penalty corresponds to a mixture prior of Gaussian and point mass at zero. Interestingly, we can show the following leads to the same local optimum set as (3.11)

$$P_{HR}(t;\mu,\eta) = \begin{cases} -\frac{1}{2}t^2 + \mu|t|, & \text{if } |t| < \frac{\mu}{1+\eta} \\ \frac{1}{2}\eta t^2 + \frac{1}{2}\frac{\mu^2}{1+\eta}, & \text{if } |t| \geq \frac{\mu}{1+\eta}. \end{cases} \tag{3.12}$$

(3.12) is continuous.

To accommodate gross outliers for the transformed model, squared-error loss is not appropriate; one can use Huber's ψ-functions from robust statistics. But a recent work of She (2011) ([50]) (IPOD) shows this amounts to adding observations shifts to (3.5) with $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$

$$b_\lambda(y_i) = b_\lambda(\sum_j x_{ij}\beta_j) + \boldsymbol{\gamma} + \varepsilon_i, \tag{3.13}$$

and then imposing sparsity on $\boldsymbol{\gamma}$. This extends the M-estimators to high-dimensional data. In summary, we will enforce a sparse prior on $\boldsymbol{\beta}$ for variable selection and another sparse prior on $\boldsymbol{\gamma}$ for outlier detection. The above algorithm can be extended to this case. We present the modiefed Θ-IPOD using an L_1 penalty on $\boldsymbol{\gamma}$ in Algorithm 2.

Algorithm 2 Modified Θ-IPOD using an L_1 penalty on $\boldsymbol{\gamma}$

Require: $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \mathbb{R}^n$, $\hat{\lambda} \in [-2,2]$, $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, $\boldsymbol{\gamma}^{(0)} \in \mathbb{R}^p$, $\mu_1 > 0$ and threshold $\Theta(.;.)$
Ensure: $j \leftarrow 0, \boldsymbol{\gamma}^{(j)} \leftarrow b_{\hat{\lambda}}(\boldsymbol{y}) - b_{\hat{\lambda}}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$, converged $\leftarrow$ FALSE
  while not converged **do**
    $\boldsymbol{r}^{(j)} \leftarrow b_{\hat{\lambda}}(\boldsymbol{y}) - b_{\hat{\lambda}}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$
    $\boldsymbol{\gamma}^{(j+1)} \leftarrow \Theta\left(\boldsymbol{r}^{(j)}; \mu_1\right)$
    converged $\leftarrow \| \boldsymbol{\gamma}^{(j+1)} - \boldsymbol{\gamma}^{(j)} \|_\infty < \varepsilon$
    $j \leftarrow j + 1$
  end while
  Return $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^{(j)}$.

We proposed the fllowing two step precedure to jointly estimate $(\hat{\lambda}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$:
1. First, we get the an initial estimation of $(\hat{\lambda}, \hat{\boldsymbol{\beta}})$ by Algorithm 1 without considering the outlier detection.
2. Then compute the mean shift parameters $\hat{\boldsymbol{\gamma}}$ via Algorithm 2 by putting L_1 penalty on $\boldsymbol{\gamma}$.
3. After getting the outliers off the observation, we go back to Algorithm 1 for updated estimator of $(\hat{\lambda}, \hat{\boldsymbol{\beta}})$ and get the final $\hat{\boldsymbol{\gamma}}$ though algorithm 2.

3.3 Simulation Study

In this section, we conduct the simulation study to evaluate the performance of proposed median regression model via TBS. The Lasso Penalized Quantile Regression (Koenker, R. 2005) ([31]) estimator is used for comparison. We report the simulation results using the following measures:
M: the mean masking probability (fraction of undetected relevant variables (*misses*))
S: the mean swamping probability (fraction of spuriously identified variables (*false alarms*))
JD: the joint successful detection rate (fraction of simulations with 0 masking)

Masking can causes more serious problem than swamping in variable selection. In an ideal scenario, we should have: **M**≈ 0%, **S**≈ 0%, and **JD**≈ 100%. Detailed description of simulation scenarios and results are summarized as follows:

· **Scenario 1** The data in simulation scenario 1 comes from the following model:

$$g_\lambda(Y_i) = g_\lambda(\mathbf{X}_i^T\beta) + \epsilon_i,\ i = 1, \ldots, n,\ , \tag{3.14}$$

Let $\lambda = 1.8$, $\beta = (3, 1.5, 0, 2, 0, 0, 0, 0)$ and ϵ_i comes from Gaussian distribution with mean=0, variance=1, $n = 200$. The correlation between X_i and X_j is $0.5^{|i-j|}$. 50 simulations were conducted here.

Table 3.1: Results for simulation study. n=200 Gaussian error, 50 replicates, 10^{-2}, sMSE=MSE$\times 10^3$

Method	Median sMSE (S.E.)	Ave. no. of non-zero	M	S	JD
TBS Lasso	1.8 (1.9)	3.3	0 %	6.8%	100%
PQ Lasso	6.8 (6.5)	4.1	0 %	38.0%	100%

· **Scenario 1 with Outliers (a)** The data in simulation scenario 1 comes from the following model:

$$g_\lambda(Y_i) = g_\lambda(\mathbf{X}_i^T\beta) + \epsilon_i,\ i = 1, \ldots, n,\ , \tag{3.15}$$

Let $\lambda = 1.8$, $\beta = (3, 1.5, 0, 2, 0, 0, 0, 0)$ and ϵ_i comes from Gaussian distribution with mean=0, variance=1, $n = 200$. The correlation between X_i and X_j is $0.5^{|i-j|}$. The shift vector is given by $\boldsymbol{\gamma} = (\{8\}^O, \{0\}^{n-O})^T$, $O \in \{5, 10, 20\}$.50 simulations were conducted here.

Table 3.2: Results for simulation study. n=200 Gaussian error, 50 replicates, 10^{-2}, sMSE=MSE$\times 10^3$

Method	Median sMSE (S.E.)	Ave. no. of non-zero	M	S	JD
TBS Lasso	2.2 (2.3)	3.3	0 %	8.4%	100%
PQ Lasso	6.5 (5.5)	4.2	0 %	38.0%	100%

· **Scenario 1 with Outliers (b)** The data in simulation scenario 1 comes from the following model:

$$g_\lambda(Y_i) = g_\lambda(\mathbf{X}_i^T\beta) + \epsilon_i,\ i = 1, \ldots, n,\ , \tag{3.16}$$

Table 3.3: Results for Outlier detection of TBS Lasso

Method	M	S	JD
TBS Lasso	0 %	2.3%	100%

Let $\lambda = 1.8$, $\beta = (3, 1.5, 0, 2, 0, 0, 0, 0)$ and ϵ_i comes from Gaussian distribution with mean=0, variance=1, $n = 200$. The correlation between X_i and X_j is $0.5^{|i-j|}$. The shift vector is given by $\boldsymbol{\gamma} = (\{8\}^O, \{0\}^{n-O})^T$, $O \in \{5, 10, 20, 25, 30, 40\}$.50 simulations were conducted here.

Table 3.4: Results for simulation study. n=200,p=8, Gaussian error, 50 replicates, 10^{-2}, sMSE=MSE$\times 10^3$

Method	Median sMSE (S.E.)	Ave. no. of non-zero	M	S	JD
TBS Lasso	2.7 (3.2)	3.4	0 %	10.0 %	100%
PQ Lasso	6.6 (5.0)	4.1	0 %	36.8%	100%

Table 3.5: Results for Outlier detection of TBS Lasso

Method	M	S	JD
TBS Lasso	0 %	2.4%	100%

· **Scenario 2** The data in simulation scenario 2 comes from the following model:

$$Y_i = \mathbf{X}_i^T \beta + \epsilon_i, \ i = 1, \ldots, n, \ , \tag{3.17}$$

Let $\beta = (3, 1.5, 0, 2, 0, 0, 0, 0)$ and ϵ_i comes from Double Exponential distribution with location=0, scale=2, $n = 200$. The correlation between X_i and X_j is $0.5^{|i-j|}$. 50 simulations were conducted here.

· **Scenario 3** Scenario 3 is the same as scenario 1, except that the number predictors is 50 instead of 8. We chose $\beta = (3, 1.5, 0, 2, \underbrace{0, \ldots, 0}_{46})$

· **Scenario 4** Scenario 4 is the same as scenario 2, except that the number predictors is 50 instead of 8. We chose $\beta = (3, 1.5, 0, 2, \underbrace{0, \ldots, 0}_{46})$

· **Scenario 5** The data in simulation scenario 5 comes from the following model: Let $\beta = (1, \underbrace{0, \ldots, 0}_{13})$, $exp(Y_i)$ comes from Exponential distribution with median = $exp(\mathbf{X}_i^T \beta)$. In the way, we have Y_i simulated from Extreme Value distribution with median = $\mathbf{X}_i^T \beta$. The correlation between X_i and X_j is $0.5^{|i-j|}$. 50 simulations were conducted here.

From the various simulation scenarios described above, We can found that TBS Lasso performed best among different simulation scenarios. The swamping of TBS Lasso is much

Table 3.6: Results for simulation study. n=200, p=8, Double-Exponential error, 50 replicates, 10^{-2}, sMSE=MSE$\times 10^2$

Method	Median sMSE (S.E.)	Ave. no. of non-zero	M	S	JD
TBS Lasso	5.5 (5.0)	3.7	0 %	14.8%	100%
PQ Lasso	4.1 (6.4)	4.2	0 %	40.4%	100%

Table 3.7: Results for simulation study. n=200, p=50, Gaussian error, 50 replicates, 10^{-2}, sMSE=MSE$\times 10^3$

Method	Median sMSE (S.E.)	Ave. no. of non-zero	M	S	JD
TBS Lasso	2.5 (2.5)	3.46	0 %	1.2%	100%
PQ Lasso	23.1 (9.8)	10.6	0 %	31.1%	100%

better while the masking is as good as PQ Lasso. The MSE of TBS Lasso are substantially smaller compared to that of PQ Lasso. The proposed method can also conduct the outlier detection simultaneously while have better performance in variable selection in median regression. All of these advantages make our proposed methods extremely attractive alternatives to other existing regularized median regression and outlier detection.

3.4 Discusssion

In this section, we developed a new regularized median regression procedure via transform-both-sides model. The proposed model can deal with heteroscedasticity of response, while successfully and jointly estimate the regression parameters and detect outliers using sparsity penalization. In this framework, we have found that our regression model work better than competing method under different simulated case. Further investigation on high dimensional problems $n \ll p$ will be conducted in the future to check whether this superior property holds.

Table 3.8: Results for simulation study. n=200, p=50, Double-Exponential error, 50 replicates, 10^{-2}, sMSE=MSE$\times 10^2$

Method	Median sMSE (S.E.)	Ave. no. of non-zero	M	S	JD
TBS Lasso	12.4 (6.4)	3.9	0 %	2.9%	100%
PQ Lasso	16.6 (10.1)	13.1	0 %	41.2%	100%

Table 3.9: Results for simulation study. n=200, p=13, Extreme Value distribution error, 50 replicates, 10^{-2}, sMSE=MSE$\times 10^2$

Method	Median sMSE (S.E.)	Ave. no. of non-zero	M	S	JD
TBS Lasso	3.2 (2.8)	1.26	0 %	3.3%	100%
PQ Lasso	4.7 (4.8)	2.9	0 %	31.7%	100%

APPENDIX A

THE PROOFS OF THEOREMS IN CHAPTER 1

Theorem 1. It will be sufficient to prove the following: If $Y = \beta x + \epsilon^*$ and $g_\lambda(Y) = g_\lambda(\beta^* x) + \epsilon^{**}$ $\epsilon^* \sim F^*$ and $\epsilon^{**} \sim F^{**}$ for some λ and two symmetric unimodal distributions F^* and F^{**} around 0, then $F^* = F^{**}$, $\beta = \beta^*$ and $\lambda = 1$.

Note that $P(Y < y|x) = F^*(y - \theta) = F^{**}\{g_\lambda(y) - g_\lambda(\theta^*)\}$ for all y, where $\theta = \beta x$ and $\theta^* = \beta^* x$. Taking $y = \theta$, we get $F^*(0) = F^{**}(0) = F^{**}\{g_\lambda(\theta) - g_\lambda(\theta^*)\} = 1/2 \Rightarrow g_\lambda(\theta) = g_\lambda(\theta^*)$. Because g_λ is monotone, this implies $\theta = \theta^*$ and the rest of the proof follows from there.

□

Remark 1. Under the TBS model, the hazard function is expressed as follows:

$$\begin{aligned} h(t|Z) &= -\frac{d}{dt}\log S_Z(t) \\ &= -\frac{d}{dt}\log\left[\Phi_0\left(\frac{g_\lambda(\log t) - g_\lambda(\beta' Z)}{\sigma}\right)\right] \\ &= \frac{\phi_0(\frac{g_\lambda(\log t) - g_\lambda(\beta' Z)}{\sigma})}{\Phi_0\left(\frac{g_\lambda(\log t) - g_\lambda(\beta' Z)}{\sigma}\right)}\frac{1}{\sigma t}(|\log t|^{\lambda-1}) \\ &= \frac{\phi_0(\omega)}{\Phi_0(\omega)}\frac{1}{\sigma t}(|\log t|^{\lambda-1}) \end{aligned}$$

where $\omega = \frac{g_\lambda(\log t) - g_\lambda(\beta' Z)}{\sigma}$, $\phi_0(\omega)$ is the standard normal $N(0,1)$ density function, $\Phi_0(\omega) = \int_\omega^{+\infty} \phi_0(u)du$ is the survival function corresponding to the density $\phi_0(\omega)$.

APPENDIX B

THE PROOFS OF THEOREMS IN CHAPTER 2

Theorem 3. Let the random variable $\mathcal{M} = \sum_{j=1}^{K} p_j U_j$, where each U_j is a random variable with uniform distribution and support in $(-\theta_j, \theta_j)$, $\theta_j > 0$, $j = 1, \ldots, K$. Take $G(\cdot)$ with discrete distribution is equivalent to take $\varepsilon_i \sim \mathcal{M}$, $i = 1, \ldots, n$. $\mathcal{M}$ is a random variable with distribution of K symmetrical uniform mixture random variables, all with mean and median zero.

For simplicity consider $K = 3$. The interval $(\widetilde{A}_i, \widetilde{B}_i]$ in Figure B.1 can be evaluate as

$$\begin{aligned}
P(\varepsilon_i \in (\widetilde{A}_i, \widetilde{B}_i]) &= P(\widetilde{A}_i < \mathcal{M} \leq \widetilde{B}_i) \\
&= P(\widetilde{A}_i < \mathcal{M} \leq -\theta_2) + P(-\theta_2 < \mathcal{M} \leq -\theta_1) \\
&\quad + P(-\theta_1 < \mathcal{M} \leq \widetilde{B}_i) \\
&= [-\theta_2 - \widetilde{A}_i]h_3 + [-\theta_1 - (-\theta_2)]h_2 + [\widetilde{B}_i - (-\theta_1)]h_1 \\
&= [-\theta_2 - \widetilde{A}_i]\left(\frac{p_3}{2\theta_3}\right) + [-\theta_1 + \theta_2]\left(\frac{p_3}{2\theta_3} + \frac{p_2}{2\theta_2}\right) \\
&\quad + [\widetilde{B}_i + \theta_1]\left(\frac{p_3}{2\theta_3} + \frac{p_2}{2\theta_2} + \frac{p_1}{2\theta_1}\right) \\
&= p_1\frac{\widetilde{B}_i + \theta_1}{2\theta_1} + p_2\frac{\widetilde{B}_i + \theta_2}{2\theta_2} + p_3\frac{\widetilde{B}_i - \widetilde{A}_i}{2\theta_3},
\end{aligned}$$

where $h_\ell = \sum_{j=\ell}^{K} \frac{p_j}{2\theta_j}$.

For $K \geq 1$ and any interval $(\widetilde{A}_i, \widetilde{B}_i]$ we have

$$P(\varepsilon_i \in (\widetilde{A}_i, \widetilde{B}_i]) = \sum_{j=1}^{K} p_j \frac{\widetilde{B}_{ij} - \widetilde{A}_{ij}}{2\theta_j},$$

where $\widetilde{A}_{ij} = \min\{\max\{-\theta_j, \widetilde{A}_i\}, \theta_j\}$ and $\widetilde{B}_{ij} = \max\{\min\{\theta_j, \widetilde{B}_i\}, -\theta_j\}$. □

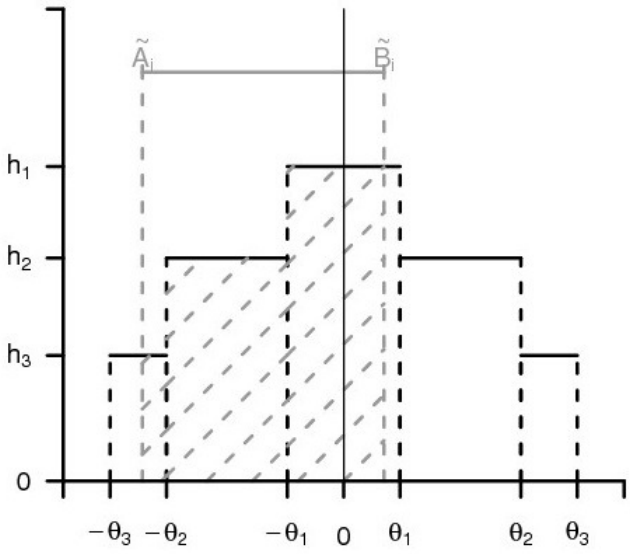

Figure B.1: Error intervals.

APPENDIX C

SOME DETAILS ON THEOREMS IN CHAPTER 3

Recall that the objective function is given by

$$f(\boldsymbol{\beta}) := l(\boldsymbol{\beta}) + P(\boldsymbol{\beta}; \mu) = \frac{1}{2}\|b_\lambda(\boldsymbol{y}) - b_\lambda(\boldsymbol{X}\boldsymbol{\beta})\|_2^2 + \sum_{j=1}^{p} P(|\beta_i|; \mu). \tag{C.1}$$

(Note that with a bit abuse of notation, we also write the penalty part as $P(\boldsymbol{\beta}'; \mu)$ or $P(|\boldsymbol{\beta}'|; \mu)$.) Simple calculation shows that for $\lambda > 0$, $b'_\lambda(t) = |t|^{\lambda-1}$ and $b''_\lambda(t) = (\lambda - 1)|t|^{\lambda-2}\mathrm{sgn}(t)$.

Definition C.0.1 (Threshold function)**.** *A threshold function is a real valued function* $\Theta(t; \mu)$ *defined for* $-\infty < t < \infty$ *and* $0 \le \mu < \infty$ *such that*

1. $\Theta(-t; \mu) = -\Theta(t; \mu)$,
2. $\Theta(t; \mu) \le \Theta(t'; \mu)$ *for* $t \le t'$,
3. $\lim_{t\to\infty} \Theta(t; \mu) = \infty$, *and*
4. $0 \le \Theta(t; \mu) \le t$ *for* $0 \le t < \infty$.

In words, $\Theta(\cdot; \mu)$ is an odd monotone unbounded shrinkage rule for t, at any μ. A vector version of Θ (still denoted by Θ) is defined componentwise if either t or μ is replaced by a vector. Given any $\mu \ge 0$, define

$$\Theta^{-1}(u; \mu) \triangleq \sup\{t : \Theta(t; \mu) \le u\}, \tag{C.2}$$

$$s(u; \mu) \triangleq \Theta^{-1}(u; \mu) - u. \tag{C.3}$$

Note that Θ^{-1} is monotonically increasing and so its derivative is defined almost everywhere on $(0, \infty)$.

C.1 Computation

Now we define Θ-estimators. Given any threshold functions Θ, the induced Θ-estimator satisfies the following nonlinear equation

$$\boldsymbol{\beta} = \Theta\left(\boldsymbol{\beta} - \boldsymbol{X}^T \mathrm{diag}\{[b'_\lambda(\boldsymbol{X}\boldsymbol{\beta})]\}(b_\lambda(\boldsymbol{X}\boldsymbol{\beta}) - b_\lambda(\boldsymbol{y})); \mu\right), \tag{C.4}$$

where b'_λ is applied in a componentwise manner and thus $\mathrm{diag}\{[b'_\lambda(\boldsymbol{X}\boldsymbol{\beta})]\} = \mathrm{diag}\{b'_\lambda(\boldsymbol{x}_i^T\boldsymbol{\beta})\}$ for $\boldsymbol{X}^T = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$.

We use $nz(\boldsymbol{\beta})$ to denote the index set of the nonzero components in β, and $z(\boldsymbol{\beta}) = (nz(\boldsymbol{\beta}))^c$. Let $D_a f$ be the directional derivative of f along direction $a \neq 0, a \in \mathbb{R}^p$. Let $\boldsymbol{e}_j = [0, \cdots, 0, 1, 0, \cdots, 0]^T$ with the jth component being 1.

Proposition 2. *Given any thresholding rule Θ, let*

$$P(\theta; \mu) - P(0; \mu) = \int_0^{|\theta|} s(u; \mu)\, \mathrm{d}u. \tag{C.5}$$

Any local minimizer $\hat{\boldsymbol{\beta}}$ of f as defined in (C.1) satisfies $D_{\boldsymbol{e}_j} f \geq 0$ and $D_{-\boldsymbol{e}_j} f \geq 0$, and thus the Θ-equation (C.4).

Proof. A useful fact is that although f is not differentiable and may be nonconvex, $D_{\pm \boldsymbol{e}_j}$ exists for any β. The first claim is from Theorem 2 of Luenberger (1997) ([36]) [P178]. Simple calculation shows

$$\boldsymbol{x}_i^T \mathrm{diag}\{[b'_\lambda(\boldsymbol{X}\boldsymbol{\beta})]\}(b_\lambda(\boldsymbol{X}\boldsymbol{\beta} - b_\lambda(\boldsymbol{y})) + s(|\beta_j|; \mu)\mathrm{sgn}(\beta_j) = 0, \forall j \in nz(\hat{\boldsymbol{\beta}}) \tag{C.6}$$

$$\boldsymbol{x}_i^T \mathrm{diag}\{[b'_\lambda(\boldsymbol{X}\boldsymbol{\beta})]\}(b_\lambda(\boldsymbol{X}\boldsymbol{\beta} - b_\lambda(\boldsymbol{y})) \in [-s(|\beta_j|; \mu), s(|\beta_j|; \mu)], \forall j \in z(\hat{\boldsymbol{\beta}}). \tag{C.7}$$

From Lemma 1 and Lemma 2 in She (2011) ([50]), $\hat{\boldsymbol{\beta}}$ is necessarily a Θ-estimate. □

So the computational problem is how to obtain a Θ-estimate satisfying (C.4). Given $\boldsymbol{\beta}^{(0)}$, the main iteration of the proposed algorithm runs as follows:

$$\boldsymbol{\beta}^{(k+1)} = \Theta\left(\boldsymbol{\beta}^{(k)} - \boldsymbol{X}^T \mathrm{diag}\{[b'_\lambda(\boldsymbol{X}\boldsymbol{\beta}^{(k)})]\}(b_\lambda(\boldsymbol{X}\boldsymbol{\beta}^{(k)}) - b_\lambda(\boldsymbol{y})); \mu\right). \tag{C.8}$$

For notational simplicity, we define

$$\boldsymbol{G}_\lambda(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{X}^T \mathrm{diag}\{[b'_\lambda(\boldsymbol{X}\boldsymbol{\beta})]\}(b_\lambda(\boldsymbol{X}\boldsymbol{\beta}) - b_\lambda(\boldsymbol{y})) \tag{C.9}$$

$$\boldsymbol{H}_\lambda(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{X}^T \mathrm{diag}\{[b''_\lambda(\boldsymbol{X}\boldsymbol{\beta})][b_\lambda(\boldsymbol{X}\boldsymbol{\beta}) - b_\lambda(\boldsymbol{y})] + [b'^2_\lambda(\boldsymbol{X}\boldsymbol{\beta})]\}\boldsymbol{X}, \tag{C.10}$$

where, according to our convention, the vector $[b''_\lambda(\boldsymbol{X}\boldsymbol{\beta})][b_\lambda(\boldsymbol{X}\boldsymbol{\beta}) - b_\lambda(\boldsymbol{y})] + [b'^2_\lambda(\boldsymbol{X}\boldsymbol{\beta})] =: d_\lambda(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y})$ is formed by elementwise operations $b''_\lambda(\boldsymbol{x}_i^T\boldsymbol{\beta})(b_\lambda(\boldsymbol{x}_i^T\boldsymbol{\beta}) - b_\lambda(y_i)) + b'^2_\lambda(\boldsymbol{x}_i^T\boldsymbol{\beta})$.

Theorem 5. *Let Θ be an arbitrarily given thresholding rule and $\boldsymbol{\beta}^{(0)}$ be any p-dimensional vector. Denote by $\boldsymbol{\beta}^{(k)}, k = 1, 2, \cdots$, the sequence of iterates obtained via (C.8). Assume Θ is continuous at any point in the closure of $\{\boldsymbol{\beta}^{(k)} - \boldsymbol{X}^T diag\{[b'_\lambda(\boldsymbol{X}\boldsymbol{\beta}^{(k)})]\}(b_\lambda(\boldsymbol{X}\boldsymbol{\beta}^{(k)}) - b_\lambda(\boldsymbol{y}))\}$. Define $\kappa = \|\boldsymbol{X}\|_2^2 \sup_{\boldsymbol{\xi} \in A} \|d_\lambda(\boldsymbol{\xi}, \boldsymbol{X}, \boldsymbol{y})\|_\infty$ where $\|\boldsymbol{X}\|_2$ is the spectral norm of $\boldsymbol{X}$, and $A = \{\vartheta\boldsymbol{\beta}^{(k)} + (1 - \vartheta)\boldsymbol{\beta}^{(k+1)} : \vartheta \in (0, 1), k = 1, 2, \cdots\}$. If $\kappa \leq 1$, then for the penalty*

functions P satisfying (C.5), the value of the corresponding objective function f decreases at each iteration

$$f(\boldsymbol{\beta}^{(k)}) - f(\boldsymbol{\beta}^{(k+1)}) \geq \frac{1-\kappa}{2}\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k+1)}\|_2^2, \quad k = 1, 2, \cdots \tag{C.11}$$

If, further, $\kappa < 1$, then any limit point of $\boldsymbol{\beta}^{(k)}$ must be a fixed point of (C.4), or a Θ-estimate.

Proof. Introduce a surrogate function

$$F(\boldsymbol{\beta}, \boldsymbol{\beta}') = l(\boldsymbol{\beta}; \lambda) + P(\boldsymbol{\beta}'; \mu) + \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 + g^T(\boldsymbol{\beta}; \lambda)(\boldsymbol{\beta}' - \boldsymbol{\beta}). \tag{C.12}$$

It follows from Lemma 1 and Lemma 2 (She 2011 ([50])) that given $\boldsymbol{\beta}$, the optimal $\boldsymbol{\beta}'_{opt} = \Theta\left(\boldsymbol{\beta} - \boldsymbol{X}^T \text{diag}\{[b'_\lambda(\boldsymbol{X}\boldsymbol{\beta})\}(b_\lambda(\boldsymbol{X}\boldsymbol{\beta}) - b_\lambda(\boldsymbol{y})); \mu\right)$. Moreover, it is not difficult to show that $F(\boldsymbol{\beta}, \boldsymbol{\beta}') \geq f(\boldsymbol{\beta}') + \frac{1-\|\boldsymbol{H}(\boldsymbol{\xi})\|_2^2}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2$. Therefore, $f(\boldsymbol{\beta}^{(k)}) - f(\boldsymbol{\beta}^{(k+1)}) \geq \frac{1-\kappa}{2}\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k+1)}\|_2^2$. The rest of the proof follows the same lines of that of Theorem 1 in She (2011) ([50]). Details omitted. □

The above theorem applies to $p > n$. In implementation, we can reduce the norm of $\boldsymbol{X}$ to guarantee the algorithm convergence from the above algorithm. That is we perform $\boldsymbol{X} \leftarrow X/K$ before running the iteration steps. Suppose n is large. Then $\boldsymbol{H}_\lambda(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}) \approx \boldsymbol{X}^T \text{diag}\{b'^2_\lambda(\boldsymbol{X}\boldsymbol{\beta}^*)\}\boldsymbol{X}$. Empirically, $K \geq \|\boldsymbol{X}\|_2\|b'_\lambda(X\hat{\boldsymbol{\beta}}_0)\|_\infty$ usually suffices for $n > p$ problems, where $\hat{\boldsymbol{\beta}}_0$ is an unpenalized estimate from minimizing $l(\boldsymbol{\beta})$.

C.2 Asymptotics

Assume

$$b_\lambda(\boldsymbol{y}) = b_\lambda(\boldsymbol{X}\boldsymbol{\beta}^*) + \boldsymbol{\varepsilon}, \quad \varepsilon_i \overset{i.i.d.}{\sim} F, \tag{C.13}$$

where F is the noise distribution. In this subsection, we assume F is $\mathcal{N}(0, \sigma^2)$.

We introduce some further notation and conditions. Let $nz^* = nz(\boldsymbol{\beta}^*)$. Suppose the following regularity conditions hold:

1. $(y_i, \boldsymbol{x}_i^T)^T$ $(1 \leq i \leq n)$ are i.i.d. following the same distribution as the random vector $(y, \boldsymbol{x}^T)^T$;
2. $\boldsymbol{\mathcal{I}}_\lambda(\boldsymbol{\beta}^*) := E\{\boldsymbol{x}\boldsymbol{x}^T b'^2_\lambda(\boldsymbol{x}^T\boldsymbol{\beta}^*)\}$ and $\boldsymbol{\mathcal{I}}_{\lambda,nz^*}(\boldsymbol{\beta}^*_{nz^*}) := E\{\boldsymbol{x}_{nz^*}\boldsymbol{x}_{nz^*}^T b'^2_\lambda(\boldsymbol{x}_{nz^*}^T\boldsymbol{\beta}^*_{nz^*})\}$ are finite and positive definite;
3. $|\frac{\partial^3 l(\boldsymbol{\beta}, y, \boldsymbol{x})}{\partial\beta_{j_1}\partial\beta_{j_2}\partial\beta_{j_3}}| \leq M_{j_1 j_2 j_3}(y, \boldsymbol{x})$ for any $\boldsymbol{\beta}$ in an open neighborhood of $\boldsymbol{\beta}^*$, and $EM_{j_1 j_2 j_3}(y, \boldsymbol{x}) < \infty$;
4. $s(\cdot; \mu_n)$ is differentiable in a neighborhood of β_j^* $\forall j \in nz^*$, say, $\{\gamma_j : |\gamma_j - \beta_j^*| \leq \delta_0, \forall j \in nz^*\}$ for some $\delta_0 > 0$, and satisfies the Lipschitz condition uniformly: $|s'(|\beta_j^* + \delta|; \mu_n) - s'(|\beta_j^*|; \mu_n)| \leq C|\delta|$, $\forall j \in nz^*$, $\forall |\delta| \leq \delta_0$ for some constant C not dependent on μ_n;

We use the same functions $\boldsymbol{G}_\lambda(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y})$ and $\boldsymbol{H}_\lambda(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y})$ as defined in (C.9) and (C.10). We also define $g_{nz^*}(\mu_n) = \max s(|\boldsymbol{\beta}^*_{nz^*}|; \mu_n)$, $h_{nz^*}(\mu_n) = \|s'(|\boldsymbol{\beta}^*_{nz^*}|; \mu_n)\|_\infty$ to control the rate of μ_n (recall that s is differentiable on $(0, \infty)$ almost everywhere).

Theorem 6. *Let $n \to \infty$. The regularization parameters are chosen such that $h_{nz^*}(\mu_n) \ll n$, $g_{nz^*}(\mu_n) = O(\sqrt{n})$, and $s(0; \mu_n) \gg \sqrt{n}$. Then there exists a sequence of Θ-estimates $\hat{\boldsymbol{\beta}}_n$ such that (i) $P(nz(\hat{\boldsymbol{\beta}}_n) = nz(\boldsymbol{\beta}^*)) \to 1$, and (ii) under the additional assumption that $s'(|\boldsymbol{\beta}^*_{nz^*}|)/n \xrightarrow{P} \boldsymbol{v}(\boldsymbol{\beta}^*_{nz^*})$ and $diag\{[s(|\boldsymbol{\beta}^*_{nz^*}|)]\} sgn(\boldsymbol{\beta}^*_{nz^*})/n \xrightarrow{P} \boldsymbol{b}(\boldsymbol{\beta}^*_{nz^*})$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{nz^*} - \boldsymbol{\beta}^*_{nz^*} + \tilde{\boldsymbol{\mathcal{I}}}^{-1}_{\lambda, nz^*}\boldsymbol{b}) \Rightarrow \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$, where $\tilde{\boldsymbol{\mathcal{I}}}_{\lambda, nz^*}(\boldsymbol{\beta}^*_{nz^*}) = \boldsymbol{\mathcal{I}}_{\lambda, nz^*} + diag\{\boldsymbol{v}\}$ and $\boldsymbol{V} = \tilde{\boldsymbol{\mathcal{I}}}^{-1}_{\lambda, nz^*}\boldsymbol{\mathcal{I}}_{\lambda, nz^*}\tilde{\boldsymbol{\mathcal{I}}}^{-1}_{\lambda, nz^*}$.*

Proof. The proof follows the lines of Fan & Li (2001) ([13]). First, we study the function difference $f(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - f(\boldsymbol{\beta}^*) = l(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - l(\boldsymbol{\beta}^*) + P(|\boldsymbol{\beta}^* + \boldsymbol{\delta}|) - P(|\boldsymbol{\beta}^*|)$ for $\boldsymbol{\delta} \to 0$.

$$\begin{aligned}
P(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - P(\boldsymbol{\beta}^*) &= \sum_{j \in z^*} P(|\beta^*_j + \delta_j|) + \sum_{j \in nz^*} P(|\beta^*_j + \delta_j|) - P(|\beta^*_j|) \\
&\geq \sum_{j \in nz^*} P(|\beta^*_j + \delta_j|) - P(|\beta^*_j|) \\
&= \sum_{j \in nz^*} s(|\beta^*_j|)\mathrm{sgn}(\beta^*_j)\delta_j + \frac{1}{2}\delta_j^2(s'(|\beta^*_j|) + o_p(1)) \\
&\geq -g_{nz^*}(\mu_n)\|\boldsymbol{\delta}_{nz^*}\|_1 - (h_{nz^*}(\mu_n)/2 + o_p(1))\|\boldsymbol{\delta}_{nz^*}\|_2^2.
\end{aligned}$$

The second equality is due to regularization condition (4).

On the other hand,

$$l(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - l(\boldsymbol{\beta}^*) = \boldsymbol{G}_\lambda^T(\boldsymbol{\beta}^*, \boldsymbol{X}, \boldsymbol{y})\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^T(\boldsymbol{H}_\lambda(\boldsymbol{\beta}^*, \boldsymbol{X}, \boldsymbol{y}) + o_p(1))\boldsymbol{\delta}$$

due to regularization condition (3). By CLT, $\sqrt{n}(\boldsymbol{G}_\lambda(\boldsymbol{\beta}^*, \boldsymbol{X}, \boldsymbol{y})/n - \boldsymbol{0}) \Rightarrow \mathcal{N}(\boldsymbol{0}, \boldsymbol{\mathcal{I}}_\lambda(\boldsymbol{\beta}^*))$. By LLN, $\boldsymbol{H}_\lambda(\boldsymbol{\beta}^*, \boldsymbol{X}, \boldsymbol{y})/n \xrightarrow{a.s.} \boldsymbol{\mathcal{I}}_\lambda(\boldsymbol{\beta}^*)$. Thus

$$l(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - l(\boldsymbol{\beta}^*) \geq O_p(\sqrt{n})\|\boldsymbol{\delta}\|_2 + (n\sigma_{\min}(\boldsymbol{\mathcal{I}}) + o_p(1))\|\boldsymbol{\delta}\|_2^2/2$$

where $\sigma_{\min}$ denotes the smallest eigenvalue.

In summary, we obtain

$$\begin{aligned}
f(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - f(\boldsymbol{\beta}^*) \geq (n\sigma_{\min}(\boldsymbol{\mathcal{I}}) - h_{nz^*}(\mu_n) + o_p(1))\|\boldsymbol{\delta}\|_2^2/2 \\
-O_p(\sqrt{n} + g_{nz^*}(\mu_n))\|\boldsymbol{\delta}\|_2. \qquad \text{(C.14)}
\end{aligned}$$

Lemma 1. *Under the regularization conditions and $h_{nz^*}(\mu_n) \ll n$ and $g_{nz^*}(\mu_n) \ll n$, there exists a sequence of Θ-estimators $\hat{\boldsymbol{\beta}}_n$ such that $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^* = O_p(\frac{1}{\sqrt{n}} + \frac{g_{nz^*}(\mu_n)}{n})$.*

Let $\gamma_n := \frac{1}{\sqrt{n}} + \frac{g_{nz^*}(\mu_n)}{n}$. Then $\gamma_n \to 0$ as $n \to \infty$. First, we show that for any $\epsilon > 0$, there exists $M > 0$ and $\hat{\boldsymbol{\beta}}_n$ such that $P(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\| \leq M\gamma_n) \geq 1 - \epsilon$ for any n. On the event $A_n(M)$ that $f(\boldsymbol{\beta}) \geq f(\boldsymbol{\beta}^*)$, $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \gamma_n\tilde{\boldsymbol{\delta}}$ for any $\|\tilde{\boldsymbol{\delta}}\|_2 = M$, there must exist a local minimum and thus a Θ-estimate $\hat{\boldsymbol{\beta}}$ (Proposition 2) that is close enough to $\boldsymbol{\beta}^*$. We only need to choose M such that $P(A_n(M)) \geq 1 - \epsilon$. From (C.14) with $\boldsymbol{\delta} = \gamma_n\tilde{\boldsymbol{\delta}}$, $f(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - f(\boldsymbol{\beta}^*) \geq$

$(n\sigma_{\min}(\boldsymbol{\mathcal{I}}) - h_{nz^*}(\mu_n) + o_p(1))\gamma_n^2 \left(\|\tilde{\boldsymbol{\delta}}\|_2^2 - (n\sigma_{\min}(\boldsymbol{\mathcal{I}}) - h_{nz^*}(\mu_n) + o_p(1))^{-1}\|\tilde{\boldsymbol{\delta}}\|_2 \right)$. Hence for $M = \|\tilde{\boldsymbol{\delta}}\|_2$ large enough, $P(A_n(M)) > 1 - \epsilon \ \forall n$. Finally, we note that by choosing $\hat{\boldsymbol{\beta}}_n$ to be the local minimum closest to $\boldsymbol{\beta}^*$ for each n (which is still a local minimum due to the continuity of f), the dependence of the estimate sequence on ϵ can be eliminated.

PROOF OF PART (I).. Based on Lemma 1, $P(nz(\hat{\boldsymbol{\beta}}_n) \supset nz^*) \to 1$. To prove (i), it remains to show for any $j \in z^*$, $P(\hat{\beta}_j = 0) \to 1$, or $P(\hat{\boldsymbol{\beta}}_{nz^*} \neq \mathbf{0}, \hat{\boldsymbol{\beta}}_{z^*} = \mathbf{0}) \to 1$ for the above sequence of Θ-estimates. From Proposition 2, it suffices to show

$$P(\pm G_j(\boldsymbol{\beta}) + s(0;\mu_n) \geq 0, \forall j \in z^*, \forall \boldsymbol{\beta} : \boldsymbol{\beta}_{z^*} = \mathbf{0}, \\ \|\boldsymbol{\beta}_{nz^*} - \boldsymbol{\beta}^*_{nz^*}\|_\infty \leq M'\gamma_n) \to 1, \quad \forall M' > 0$$

where $G_j(\boldsymbol{\beta})$ is the jth component of $\boldsymbol{G}_\lambda(\boldsymbol{\beta})$. We will show $|G_j(\boldsymbol{\beta})| \ll s(0;\mu_n)$ bounded in probability for such $\boldsymbol{\beta}$ under $s(0;\mu_n) \gg \sqrt{n}$ and $g_{nz^*}(\mu_n) = O(\sqrt{n})$ (and thus $\gamma_n = 1/\sqrt{n}$).

From the regularity condition (4) and Lemma 1, for any $\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty \leq M'\gamma_n$,

$$\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{G}(\boldsymbol{\beta}^*) + (\boldsymbol{H}(\boldsymbol{\beta}^*) + o_p(1))(\boldsymbol{\beta} - \boldsymbol{\beta}^*).$$

As discussed before, the first term is $O_p(\sqrt{n})$, and $\boldsymbol{H}(\boldsymbol{\beta}^*) = O_p(n)$ and $\boldsymbol{G}(\boldsymbol{\beta}) = O_p(\sqrt{n} + n\gamma_n)$. By the assumption on $s(0;\mu_n)$, the selection consistency must hold.

PROOF OF PART (II).. Seen from the Θ-equation, $\hat{\boldsymbol{\beta}}_{nz(\hat{\boldsymbol{\beta}})}$ satisfies $G_j(\hat{\boldsymbol{\beta}}, \boldsymbol{X}, \boldsymbol{y}) + s(|\hat{\beta}_j|)\text{sgn}(\hat{\beta}_j) = 0, \forall j \in nz(\hat{\boldsymbol{\beta}})$. With probability tending to 1, $\hat{\boldsymbol{\beta}}_{nz^*}$ satisfies

$$G_j(\hat{\boldsymbol{\beta}}, \boldsymbol{X}, \boldsymbol{y}) + s(|\hat{\beta}_j|)\text{sgn}(\hat{\beta}_j) = 0, \forall j \in nz^*. \tag{C.15}$$

Define

$$\boldsymbol{G}_{nz^*}(\boldsymbol{\beta}_{nz^*}, \boldsymbol{X}_{nz^*}, \boldsymbol{y}) = \boldsymbol{X}^T_{nz^*}\text{diag}\{[b'_\lambda(\boldsymbol{X}_{nz^*}\boldsymbol{\beta}_{nz^*})]\}(b_\lambda(\boldsymbol{X}_{nz^*}\boldsymbol{\beta}_{nz^*}) - b_\lambda(\boldsymbol{y})) \\ \boldsymbol{H}_{nz^*}(\boldsymbol{\beta}_{nz^*}, \boldsymbol{X}_{nz^*}, \boldsymbol{y}) = \boldsymbol{X}^T_{nz^*}\text{diag}\{[b''_\lambda(\boldsymbol{X}_{nz^*}\boldsymbol{\beta}_{nz^*})][(b_\lambda(\boldsymbol{X}_{nz^*}\boldsymbol{\beta}_{nz^*}) - b_\lambda(\boldsymbol{y}))] \\ + [b'^2_\lambda(\boldsymbol{X}_{nz^*}\boldsymbol{\beta}_{nz^*})]\}\boldsymbol{X}_{nz^*}.$$

Again, Taylor expansions give $\boldsymbol{G}_{nz^*}(\boldsymbol{\beta}_{nz^*}, \boldsymbol{X}_{nz^*}, \boldsymbol{y}) = \boldsymbol{G}_{nz^*}(\boldsymbol{\beta}^*_{nz^*}, \boldsymbol{X}_{nz^*}, \boldsymbol{y}) + (\boldsymbol{H}_{nz^*}(\boldsymbol{\beta}_{nz^*}, \boldsymbol{X}_{nz^*}, \boldsymbol{y}) + o_p(1))(\boldsymbol{\beta}_{nz^*} - \boldsymbol{\beta}^*_{nz^*})$, and $s(|\hat{\beta}_j|)\text{sgn}(\hat{\beta}_j) = s(|\beta^*_j|)\text{sgn}(\beta^*_j) + (s'(|\beta^*_j|) + o_p(1))(\hat{\beta}_j - \beta^*_j) \ \forall j \in nz^*$, for $\boldsymbol{\beta}_{nz^*}$ close to $\boldsymbol{\beta}^*_{nz^*}$.

Plugging them into (C.15) yields for $\text{sgn}(\boldsymbol{\beta}_{nz^*}) = \text{sgn}(\hat{\boldsymbol{\beta}}_{nz^*})$

$$\boldsymbol{G}_{nz^*}(\boldsymbol{\beta}^*_{nz^*}, \boldsymbol{X}_{nz^*}, \boldsymbol{y}) + (\boldsymbol{H}_{nz^*}(\hat{\boldsymbol{\beta}}_{nz^*}, \boldsymbol{X}_{nz^*}, \boldsymbol{y}) + o_p(1))(\hat{\boldsymbol{\beta}}_{nz^*} - \boldsymbol{\beta}^*_{nz^*}) \\ + \text{diag}\{[s(|\boldsymbol{\beta}^*_{nz^*}|)]\}\text{sgn}(\boldsymbol{\beta}^*_{nz^*}) + (\text{diag}\{[s'(|\boldsymbol{\beta}^*_{nz^*}|)]\} + o_p(1))(\hat{\boldsymbol{\beta}}_{nz^*} - \boldsymbol{\beta}^*_{nz^*}) = \mathbf{0},$$

and so

$$\sqrt{n}\tilde{\boldsymbol{H}}\hat{\boldsymbol{\beta}}_{nz^*} - \boldsymbol{\beta}^*_{nz^*} + \tilde{\boldsymbol{H}}^{-1}\frac{\text{diag}\{[s(|\boldsymbol{\beta}^*_{nz^*}|)]\}\text{sgn}(\boldsymbol{\beta}^*_{nz^*})}{n} \\ = -\sqrt{n}\frac{\boldsymbol{G}_{nz^*}(\boldsymbol{\beta}^*_{nz^*}, \boldsymbol{X}_{nz^*}, \boldsymbol{y})}{n} + o_p(1).$$

where $\tilde{\boldsymbol{H}} = \{\frac{\boldsymbol{H}_{nz^*}(\hat{\boldsymbol{\beta}}_{nz^*}, \boldsymbol{X}_{nz^*}, \boldsymbol{y})}{n} + \frac{\text{diag}\{[s'(|\boldsymbol{\beta}^*_{nz^*}|)]\}}{n}\} \xrightarrow{P} \boldsymbol{\mathcal{I}}_{\lambda, nz^*}(\boldsymbol{\beta}^*) + \text{diag}\{\boldsymbol{v}(\boldsymbol{\beta}^*_{nz^*})\}$. Part (ii) follows from CLT and Slutsky's theorem. □

C.3 Robustness

Even after the simultaneous Box-Cox transforms with respect to both $\boldsymbol{y}$ and $\boldsymbol{X\beta}$, the model (C.13) may not be Gaussian. In robust statistics, instead of using the squared error loss, a robust loss function ρ is usually applied:

$$\min \sum_i \rho(b_\lambda(y_i) - b_\lambda(\boldsymbol{x}_i^T\boldsymbol{\beta})). \tag{C.16}$$

Let $\psi = \rho'$. The associated estimator, referred to as the ψ-estimator ([25]), has good robustness and efficiency. The popular choices of ψ include Huber's ψ, Hampel's three-parts ψ, Tukey's bisquare ψ, etc.

Motivated by She & Owen (2011) ([50]), we propose to keep the squared error loss function, but introduce an additional parameter $\boldsymbol{\gamma} \in \mathbb{R}^p$, and minimize

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\gamma}\in\mathbb{R}^n} \frac{1}{2}\|b_\lambda(\boldsymbol{y}) - b_\lambda(\boldsymbol{X\beta}) - \boldsymbol{\gamma}\|_2^2 + \tilde{P}(\boldsymbol{\gamma};\nu). \tag{C.17}$$

Our algorithm can be adapted to solve the problem. The main iteration step becomes

$$\begin{cases} \boldsymbol{\beta}^{(k+1)} & = \boldsymbol{\beta}^{(k)} - \boldsymbol{X}^T \mathrm{diag}\{[b'_\lambda(\boldsymbol{X\beta}^{(k)})]\}\{b_\lambda(\boldsymbol{X\beta}^{(k)}) - b_\lambda(\boldsymbol{y}) + \boldsymbol{\gamma}^{(k)}\} \\ \boldsymbol{\gamma}^{(k+1)} & = \tilde{\Theta}(b_\lambda(\boldsymbol{y}) - b_\lambda(\boldsymbol{X\beta}^{(k)});\nu) \end{cases}$$

With $\boldsymbol{X}$ sufficiently small, the algorithm converges to $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ which is referred to as the $\tilde{\Theta}$-estimator. The following result shows $\hat{\boldsymbol{\beta}}$ obtained for this additive mean shift model is a robust estimator.

Theorem 7. *A $\tilde{\Theta}$-estimator $\hat{\boldsymbol{\beta}}$ to (C.17) is a robust ψ-estimator to (C.16) provided $\tilde{\Theta}+\psi = Id$.*

Proof. $\hat{\boldsymbol{\beta}}$ satisfies

$$\begin{aligned} 0 =& \boldsymbol{X}^T \mathrm{diag}\{[b'_\lambda(\boldsymbol{X\beta})]\}(b_\lambda(\boldsymbol{X\beta}) - b_\lambda(\boldsymbol{y}) + \boldsymbol{\gamma}) \\ =& \boldsymbol{X}^T \mathrm{diag}\{[b_\lambda(\boldsymbol{X\beta})]\}(b_\lambda(\boldsymbol{X\beta}) - b_\lambda(\boldsymbol{y}) + \tilde{\Theta}(b_\lambda(\boldsymbol{y}) - b_\lambda(\boldsymbol{X\beta})) \\ =& -\boldsymbol{X}^T \mathrm{diag}\{[b_\lambda(\boldsymbol{X\beta})]\}\psi(b_\lambda(\boldsymbol{y}) - b_\lambda(\boldsymbol{X\beta})). \end{aligned}$$

□

Therefore, the breakdown-point properties and the efficiency results carry over to $\tilde{\Theta}$ estimators. For example, if we use the soft-thresholding, the $\tilde{P}$ in (C.17) is the l_1-penalty, and the loss function ρ in (C.16) is the well known Huber's robust loss function.

In our problems, $\boldsymbol{\beta}$ can be sparse, and $p > n$ is possible. Therefore, we study the robust Box-Cox variable selection selection problem defined by

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\gamma}\in\mathbb{R}^n} \frac{1}{2}\|b_\lambda(\boldsymbol{y}) - b_\lambda(\boldsymbol{X\beta}) - \boldsymbol{\gamma}\|_2^2 + P(\boldsymbol{\beta};\mu) + \tilde{P}(\boldsymbol{\gamma};\nu). \tag{C.18}$$

Our algorithm becomes

$$\begin{cases} \boldsymbol{\beta}^{(k+1)} & = \Theta\left(\boldsymbol{\beta}^{(k)} - \boldsymbol{X}^T \mathrm{diag}\{[b'_\lambda(\boldsymbol{X}\boldsymbol{\beta}^{(k)})]\}\{b_\lambda(\boldsymbol{X}\boldsymbol{\beta}^{(k)}) - b_\lambda(\boldsymbol{y}) + \boldsymbol{\gamma}^{(k)}\}; \mu\right) \\ \boldsymbol{\gamma}^{(k+1)} & = \tilde{\Theta}(b_\lambda(\boldsymbol{y}) - b_\lambda(\boldsymbol{X}\boldsymbol{\beta}^{(k)}); \nu). \end{cases}$$

Similar to Theorem 5, the convergence is guaranteed if we prescale $\boldsymbol{X}$ by $\boldsymbol{X} \leftarrow \boldsymbol{X}/K$ for large enough K.

REFERENCES

[1] Ørnulf B. Gill R. Keiding N. Andersen, P. *Statistical Models Based on Countig Processes*. New York: Spinger-Verlag., 1992.

[2] Tsiatis A. A. Bang, H. Median regression with censored cost data. *Biometrics*, 58:643–649, 2003.

[3] Doksum K.A. Bickel, P.J. An analysis of transformations revisited. *J. Amer. Statist. Assoc.*, 76:296–311, 1981.

[4] Cox D.R. Box, G.E.P. An analysis of transformations revisited. *Journal of the Royal Statistical Society, Series B*, 26:211–243, 1964.

[5] Ruppert D. Carroll, R.J. Power-transformations when fitting theoretical models to data. *J. Amer. Statist. Assoc.*, 79:321–328, 1984.

[6] Jin Z. Chen, K. and Ying. Z. Semiparametric analysis of transformation models with censored data. *Biometrika*, 89:659–668, 2002.

[7] Wei L.J. Ying Z. Cheng, S.C. Analysis of transformation models with censored data. *Biometrika*, 82:835–845, 1995.

[8] Wei L.J. Ying Z. Cheng, S.C. Predicting survival probabilities with semiparametric transformation models. *J. Amer. Statist. Assoc.*, 92:227–235, 1997.

[9] R.L. Comis. Clinical trials of cyclophosphamide, etoposide and vincristine in the treatment of small-cell lung cancer. *Semin Oncol*, 13:40–44, 1986.

[10] D. R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–200, 1997.

[11] Hastie T. Johnstone I. Efron, B. and R. Tibshirani. Least angle regression (with discussion). *The Annals of Statistics*, 32:407–451, 2004.

[12] Feld R. Murray N. et al Evans, W.K. Superiority of alternating non ross resistant chemotherapy in extensive small cell lung cancer. *Ann Intern Med*, 107:451–458, 1987.

[13] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360, 2001.

[14] W. Feller. *An Introduction to Probability Theory and Its Applications*. Wiley., 1971.

[15] T.S. Ferguson. Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

[16] Ying Z. Wei L.J. Fine, J.P. On the linear transformation model with censored data. *Biometrika*, 85:980–986, 1998.

[17] D. M. Finkelstein. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, 41:933–945, 1985.

[18] D. M. Finkelstein. A proportional hazards model for interval-censored failure time data. *Biometrics*, 42:845–854, 1986.

[19] Lipsitz S.R. Parzen M. Fitzmaurice, G.M. Approximate median regression via the box-cox transformation. *The American Statistician*, 61:233–238, 2007.

[20] Dey D. Chang H. Gelfand, A. *Model determination using predictive distributions with implementation via sampling-based methods.* In *Bayesian Statistics 4*, J. Bernardo, J. Berger, A. Dawid & A. Smith, eds. Oxford University Press., 1992.

[21] Sinha D. Ghosh, S. Bayesian analysis of interval-censored survival data using penalized likelihood. *Sankhya, Ser. A.*, 63:1–14, 2000.

[22] Johnson W. Hanson, T. A bayesian semiparametric aft model for interval-censored data. *Journal of Computational and Graphical Statistics*, 13:341–361, 2004.

[23] Johnson W.O. Hanson, T. Modeling regression error with a mixture of polya trees. *J. Amer. Statist. Assoc.*, 97:1020–1033, 2002.

[24] Yang M. Hanson, T. Bayesian semiparametric proportional odds models. *Biometrics*, 63:88–95, 2007.

[25] P. J. Huber. *Robust Statistics*. John Wiley, 1981.

[26] Chen M.-H. Sinha D. Ibrahim, J.G. *Bayesian Survival Analysis*. Springer-Verlag., 2001.

[27] Everson L.-Therneau T.M. Jett, J.R. Treatment of limited-stage small-cell lung cancer with cyclophosphamide, doxorubicin and vincristine with or without etoposide: A randomized trial of the north central cancer treatment group. *J Clin Oncol*, 8:33–38, 1990.

[28] van der Lann M. J. Jewell, N. P. *Advances in Survival Analysis, vol. 23 of Handbook of Statistics, chap. Current status data: review, recent developments and open problems.* Elsevier, 2004.

[29] S. Kettl. Accounting for heteroscedasticity in the transform both sides regression model. *Applied statistics*, 49:261–268, 1991.

[30] A.Y. Khintchine. On unimodal distributions. *Inst. Mat. Mech. Tomsk. Gos. Univ.*, 2:1–7, 1938.

[31] R. Koenker. Censored quantile regression redux. *J. Statistical Software*, 27:1–24, 2008.

[32] Cai J. Kong, L. and P. K. Sen. Weighted estimating equations for semiparametric transformation models with censored data from a casecohort design. *Biometrika*, 91:305–319, 2004.

[33] Gelfand A.E. Kottas, A. Bayesian semiparametric median regression modeling. *J. Amer. Statist. Assoc.*, 456:1458–1468, 2001.

[34] Mallick B.K. Kuo, L. Bayesian semiparametric inference for the accelerated failure-time model. *The Canadian Journal of Statistics*, 25:457–472, 1997.

[35] W. Lu and Z. Ying. On semiparametric transformation cure models. *Biometrika*, 91:331–343, 2004.

[36] D. Luenberger. *Optimization by vector space methods*. John Wiley, 1997.

[37] Thomas A. Best N. Lunn, D. J. Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337, 2000.

[38] Subramanian S. Sun Y. McKeague, I.W. Median regression and the missing information principle. *J. Nonparametric Statist.*, 13:709–727, 2001.

[39] van der Vaart A. W. Murphy, S. A. On profile likelihood. *Journal of the American Statistical Association*, 95:449–465, 2001.

[40] Mead R. Nelder, J. A. A simplex algorithm for function minimization. *Computer Journal*, 7:308–313, 1965.

[41] Branden K. V. Portnoy S. Neocleous, T. Correction to censored regression quantiles by s. portnoy, 98. *Journal of the American Statistical Association*, 101:860–861, 2006.

[42] Presnell B. Osborne, M. R. and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404, 2000.

[43] W. Pan. A multiple imputation approach to cox regression with interval-censored data. *Biometrics*, 56:199–203, 2000.

[44] Huang Y. Peng, L. Survival analysis with quantile regression models. *J. Amer. Statist. Assoc.*, 103:637–649, 2008.

[45] R. Peto. An experimental survival curve for interval-censored data. *Journal of the Royal Statistical Society, Ser. C (Applied Statistics)*, 22:86–91, 1973.

[46] S. Piantadosi. *Clinical Trials: A Methodologic Perspective*. Wiley series in probability and statistics. Wiley-Interscience, 2nd ed., 2005.

[47] S. Portnoy. Censored regression quantiles. *J. Amer. Statist. Assoc.*, 98:1001–1012, 2003.

[48] Datta S. Williamson J. Satten, G. Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association*, 93:318–327, 1994.

[49] J. Sethuraman. Constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[50] Owen A. She, Y. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106:626–639, 2011.

[51] Chen M.-H. Ghosh S. Sinha, D. Bayesian anaysis and predictive model diagnostics for interval-censored survival data. *Biometrics*, 55:585–590, 1999.

[52] Johnston G.-Kim H. So, Y. Analyzing interval-censored data with sas software. *Proceedings of the SAS Global Forum 2010 Conference*, paper 257, 2010.

[53] J. Sun. *The Statistical Analysis of Interval-Censored Failure Time Data*. No. XVI in Statistics for Biology and Health. New York: Springer., 2006.

[54] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B*, 58:267–288, 1996.

[55] Mallick B.K. Walker, S. A bayesian semiparametric accelerated failure time model. *Biometrics*, 55:477–483, 1999.

[56] H. White. Maximum likelihood estimation under misspecified models. *Econometrica*, 50:1–26, 1999.

[57] Severini-T. A. Wong, W. H. On maximum likelihood estimation in infinite dimensional parameter spaces. *The Annals of Statistics*, 19:603–632, 1991.

[58] Prentice R. Yang, S. Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association*, 94:125–136, 1999.

[59] Jung S. H. Wei L. J. Ying, Z. Survival analysis with median regression models. *J. Amer. Statist. Assoc.*, 90:178–184, 1995.

BIOGRAPHICAL SKETCH

Jianchang Lin was born in Fujian, China in Dec. 1983. In the summer of 2005, he completed the Bachelor of Science degree in Mathematical Statistics (minor in Finance) at University of Science and Technology of China (USTC). In the fall of 2007, he was admitted to the doctoral program of Statistics the Florida State University. He defended his dissertation in the spring of 2012. His current research interests include Bayesian biostatistics, Survival analysis and adaptive dose-finding design in oncology.