BAYESIAN METHODS FOR SKEWED RESPONSE INCLUDING LONGITUDINAL AND HETEROSEDASTIC DATA

By

YUANYUAN TANG

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

Degree Awarded:
Summer Semester, 2013
Yuanyuan Tang defended this dissertation on June 19, 2013.
The members of the supervisory committee were:

Debajyoti Sinha
Professor Directing Thesis

Debdeep Pati
Professor Directing Thesis

Heather Flynn
University Representative

Yiyuan She
Committee Member

Stuart Lipsitz
Committee Member

Eric Chicken
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.
This dissertation is dedicated to my husband Liangrui, my grandparents and parents.
ACKNOWLEDGMENTS

My first acknowledgement must go to my major advisors, Dr. Debajyoti Sinha and Dr. Debdeep Pati. I would like to take this opportunity to express my special thanks to them for inspiring me to study Bayesian Biostatistics, and for all of their invaluable guidance, advice and support on my dissertation research. I greatly appreciate the time and faith that they invested in me as a student.

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

Meanwhile, I am grateful to all the faculty, especially Dr. Xufeng Niu, Dr. Dan McGee, Dr. Fred Huffer, and all the staff in the Department of Statistics for their generous help during my graduate study time.

I am also thankful for my friends in the Department of Statistics. Finally, I would like to thank my husband for his support and encouragement throughout this process.
# TABLE OF CONTENTS

List of Tables ................................................................. vii
List of Figures ............................................................... viii
Abstract .............................................................. ix

1 Bayesian Partial Linear Model for Skewed Longitudinal Data 1
   1.1 Introduction ....................................................... 1
   1.2 Model Specifications ........................................... 3
   1.3 Properties of Prior and Posterior Consistency ............... 7
      1.3.1 Large support of the prior for the multivariate residual density 7
      1.3.2 Spline approximation property .......................... 11
      1.3.3 Posterior consistency .................................. 13
   1.4 Simulation Study .............................................. 15
   1.5 Analysis of P2C2 Study ....................................... 18
   1.6 Conclusions ..................................................... 21

2 Bayesian Variable Selection for Skewed Heteroscedastic Error 23
   2.1 Introduction ..................................................... 23
   2.2 Transform-both-sides Model and Prior ....................... 25
      2.2.1 Transform-both-sides model ........................... 25
      2.2.2 Outlier or heavy tail distribution ..................... 27
   2.3 Data Example .................................................... 29
   2.4 Simulation Studies ............................................. 38
      2.4.1 None-outlier ........................................... 38
      2.4.2 Heavy tail distribution ............................... 40
   2.5 Discussion ...................................................... 42

3 Practical Semiparametric Bayes Analysis of Heteroscedastic and Skewed Response 43
   3.1 Introduction ..................................................... 43
   3.2 Model and Prior Specification ............................... 45
      3.2.1 Model for heteroscedastic residual density ......... 45
      3.2.2 Choice of $\Pi_{\mu}$ .................................. 46
      3.2.3 Prior flexibility ....................................... 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.4 Analogy with discrete mixture models</td>
<td>49</td>
</tr>
<tr>
<td>3.3 Posterior Computation</td>
<td>49</td>
</tr>
<tr>
<td>3.4 Simulation Studies</td>
<td>51</td>
</tr>
<tr>
<td>3.5 An Epidemiological Application</td>
<td>51</td>
</tr>
<tr>
<td>3.6 Discussion</td>
<td>54</td>
</tr>
<tr>
<td>A The proof of theorems in chapter 1</td>
<td>58</td>
</tr>
<tr>
<td>A.1 Proof of Equation 1.4</td>
<td>58</td>
</tr>
<tr>
<td>A.2 Proof of correlation</td>
<td>58</td>
</tr>
<tr>
<td>B Proof for chapter 3</td>
<td>60</td>
</tr>
<tr>
<td>B.1 prior for $\gamma$</td>
<td>60</td>
</tr>
<tr>
<td>C The proof of theorems in chapter 2</td>
<td>61</td>
</tr>
<tr>
<td>C.1 Proof of Proposition 1</td>
<td>61</td>
</tr>
<tr>
<td>References</td>
<td>63</td>
</tr>
<tr>
<td>Biographical Sketch</td>
<td>68</td>
</tr>
</tbody>
</table>
LIST OF TABLES

1.1 Approximate (via Monte Carlo) sampling mean and mean square error (MSE within parenthesis) of the estimated regression parameters ($\beta_1, \beta_2$); and approximate mean integrated squared error (MISE) for function $v(t)$. True parameter values are $\beta_1 = 3, \beta_2 = -1$. . . . . . . . 17

1.2 Results of the analysis of HIV study with the estimators and the corresponding standard errors (within parenthesis) of the regression parameters ($\beta_1, \beta_2$) and other model parameters under different methods. . . 21

2.1 Results for simulation study: Scenario 1. . . . . . . . . . . . . . . . . . . 39

2.2 Results for simulation study: Scenario 2. . . . . . . . . . . . . . . . . . . 39

2.3 Comparison of Spike Gaussian model and Penalized Quantile model. . 41

2.4 Results for simulation study: Scenario 3. . . . . . . . . . . . . . . . . . . 42

2.5 Results for simulation study: Scenario 4. . . . . . . . . . . . . . . . . . . 42

3.1 Approximate (via Monte Carlo) sampling mean and mean square error (1000×MSE within parenthesis) of the estimated regression parameters ($\beta$); True parameter values are $\beta = 2$. . . . . . . . . . . . . . . . . . 52

3.2 Approximate (via Monte Carlo) sampling mean and mean square error (1000×MSE within parenthesis) of the estimated regression parameters ($\beta_1, \beta_2, \beta_3$); True parameter values are $\beta_1 = -3, \beta_2 = 2, \beta_3 = 5$. . . . . . 52
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Variogram Plot: HIV=0</td>
<td>19</td>
</tr>
<tr>
<td>1.2</td>
<td>Variogram Plot: HIV=1</td>
<td>20</td>
</tr>
<tr>
<td>2.1</td>
<td>Histogram of Medical Cost for Training Data</td>
<td>30</td>
</tr>
<tr>
<td>2.2</td>
<td>QQ plot and Histogram for the $U^T_1 y$ in the training data</td>
<td>31</td>
</tr>
<tr>
<td>2.3</td>
<td>QQ plot and Histogram for Loss2 and Loss3: Penalized Quantile</td>
<td>32</td>
</tr>
<tr>
<td>2.4</td>
<td>QQ plot and Histogram for Loss2 and Loss3: TBS</td>
<td>33</td>
</tr>
<tr>
<td>2.5</td>
<td>Histogram of Medical Cost for Test Data</td>
<td>34</td>
</tr>
<tr>
<td>2.6</td>
<td>Empirical CDF 95%</td>
<td>35</td>
</tr>
<tr>
<td>2.7</td>
<td>Empirical CDF 99.5%</td>
<td>36</td>
</tr>
<tr>
<td>2.8</td>
<td>Empirical CDF 100%</td>
<td>37</td>
</tr>
<tr>
<td>3.1</td>
<td>Density Plot of Skewed and Heteroscedastic Error</td>
<td>47</td>
</tr>
<tr>
<td>3.2</td>
<td>Histogram of GAD</td>
<td>53</td>
</tr>
<tr>
<td>3.3</td>
<td>Plot of the log-ratio of CPOs obtained from LV-RD and Gaussian, from LV-RD and SDE model (y-axis), versus DDE (x-axis).</td>
<td>54</td>
</tr>
<tr>
<td>3.4</td>
<td>Predicted quantiles versus DDE: Gaussian model</td>
<td>55</td>
</tr>
<tr>
<td>3.5</td>
<td>Predicted quantiles versus DDE: SDE model</td>
<td>56</td>
</tr>
<tr>
<td>3.6</td>
<td>Predicted quantiles versus DDE: LV-RD model</td>
<td>57</td>
</tr>
</tbody>
</table>
ABSTRACT

Skewed response data are very popular in practice, especially in biomedical area. We begin our work from the skewed longitudinal response without heteroscedasticity. We extend the skewed error density to the multivariate response. Then we study the heteroscedasticity. We extend the transform-both-sides model to the bayesian variable selection area to handle the univariate skewed response, where the variance of response is a function of the median. At last, we proposed a novel model to handle the skewed univariate response with a flexible heteroscedasticity.

For longitudinal studies with heavily skewed continuous response, statistical model and methods focusing on mean response are not appropriate. In this paper, we present a partial linear model of median regression function of skewed longitudinal response. We develop a semiparametric Bayesian estimation procedure using an appropriate Dirichlet process mixture prior for the skewed error distribution. We provide justifications for using our methods including theoretical investigation of the support of the prior, asymptotic properties of the posterior and also simulation studies of finite sample properties. Ease of implementation and advantages of our model and method compared to existing methods are illustrated via analysis of a cardiotoxicity study of children of HIV infected mother.

Our second aim is to develop a Bayesian simultaneous variable selection and estimation of median regression for skewed response variable. Our hierarchical Bayesian model can incorporate advantages of $l_0$ penalty for skewed and heteroscedastic error. Some preliminary simulation studies have been conducted to compare the performance of proposed model and existing frequentist median lasso regression model. Considering the estimation bias and total square error, our proposed model performs as good as, or better than competing frequentist estimators.

In biomedical studies, the covariates often affect the location, scale as well as the shape of the skewed response distribution. Existing biostatistical literature mainly focuses on the mean regression with a symmetric error distribution. While such modeling assumptions and methods are often deemed as restrictive and inappropriate for skewed response, the completely nonparametric methods may lack a physical interpretation of the covariate effects. Existing nonparametric methods also miss any easily implementable computational tool. For a skewed response, we develop a novel model accommodating a nonparametric error density that depends on the covariates. The advantages of our semiparametric associated Bayes method include the ease of prior elicitation/determination, an easily implementable posterior computation,
theoretically sound properties of the selection of priors and accommodation of possible outliers. The practical advantages of the method are illustrated via a simulation study and an analysis of a real-life epidemiological study on the serum response to DDT exposure during gestation period.
CHAPTER 1

BAYESIAN PARTIAL LINEAR MODEL FOR SKEWED LONGITUDINAL DATA

1.1 Introduction

Existing methods for analysis of longitudinal continuous response including, for example, the mixed model framework (Diggle et al., 2002) e.g. with identity link, generally focus on estimating the mean response function when the outcome variable is either approximately Gaussian or symmetric. However, the assumption of symmetric errors may be inappropriate for the analysis of many real-life biomedical, epidemiological and clinical studies. For example, in a Pediatric Pulmonary and Cardiac Complications (in short P2C2) study of vertically transmitted HIV infection to children of infected mothers, the longitudinal response of interest RRBO (interval between consecutive R-waves measured via EKG, e.g., the time between heartbeats) is known to be heavily skewed. The EKG is often taken in a cardiologist’s office, and the RRBO is used to screen for cardiac problems. For example, if the RRBO is too short (less than 0.6 seconds), then the patient’s heart rate (pulse) is too rapid (tachycardia) and the patient could have a serious cardiac problem such as atrial fibrillation. If the RRBO is too long (more than 1 second), then the heart rate is too slow (bradycardia) and the patient could also have a serious cardiac problem such as ischemic heart disease. In general, it is difficult to determine a suitable transformation of the RRBO to achieve the assumption of symmetric error. There is also concern about possible interpretation of covariate effects on the actual response. It is also difficult to specify and elicit prior opinions about model parameters and functions of the transformed model based on the available prior opinions of the untransformed response variable. This is particularly a critical practical issue for Bayesian statistical analysis when the transformation is either unknown or have unknown parameter. For a skewed response, a model based on median regression is an attractive alternative to mean regression model. He et al. (2003) present an in-depth review and comparison of linear median regression models for longitudinal data using various frequentist marginal analysis approaches including those by Koenker and Bassett Jr (1978), Jung (1996), and Huggins (1993). However, all these methods assume parametric regres-
sion functions. These methods also do not specify the likelihood, and hence are not suitable for prediction with proper assessment of uncertainty. Kottas and Gelfand (2001), Hanson and Johnson (2002), and Lin et al. (2011) present Bayesian median regression models for a skewed univariate response. However, these existing semi-parametric Bayes methods for median regression do not deal with either multivariate or longitudinal data. We also present a class of skewed marginal density for the error along with a class of prior process based on Dirichlet mixture of uniforms. We demonstrate a major practical advantage of our prior process — the ease of the specification of the process based on available prior opinion.

Partial linear models have been used extensively due to their ability to incorporate nonparametric function for the effects of some of the covariates, while maintaining the physical interpretation of the parametric regression function for the rest of the covariates of interest. The partial linear model is especially appropriate for longitudinal studies as the effect of time on response is often an unknown nonlinear function. For such studies including our P2C2 study, it is desirable to model the time-effect as a nonparametric function whereas using a parametric regression parameter for the covariate of interest (say, HIV status) to allow good and simple physical/clinical interpretation. Speckman (1988) studies kernel smoothing in partial linear models with uncorrelated mean zero errors. Liang et al. (1999) introduce a partial linear model with an unknown function for a scalar covariate with independent mean zero errors. He et al. (2002) extend the \( M \)-estimators of univariate semiparametric models to the case of longitudinal data using a regression spline for the nonparametric function. The form of the error distribution and the dependence structure within each subject are not specified. We show here that our likelihood based Bayes analysis, not unexpectedly, provides a more precise estimator compared to the estimator based on \( M \)-estimation method. However, to the best of our knowledge, no one has yet developed a model and likelihood based analysis for a partial linear model for the median with heavily skewed longitudinal responses.

In this paper, we provide a Bayes method of analysis with a good physical interpretation, ease of implementation and complete theoretical justification. In particular, we show strong posterior consistency in estimating the regression coefficients without requiring restrictive modeling and regularity assumptions. Although consistency of frequentist estimators has been widely studied in a partial linear model involving both parametric and nonparametric error densities (Härdle et al., 2000; Bhattacharya and Zhao, 1997), analogous results in a fully Bayesian paradigm are lacking with the exception of Bickel and Kleijn (2012), for a partial linear model, however, only with a parametric error distribution.

In a Bayesian paradigm, consistency in semiparametric regression models with univariate response has been studied by Amewou-Atisso et al. (2003), Pati and Dunson (2012) and very recently by Pelenis (2012). While Pati and Dunson (2012) and Amewou-Atisso et al. (2003) consider estimation of mean regression with symmetric residual densities, Pelenis (2012), in an unpublished manuscript, deals with mean regression with asymmetric heteroscedastic error. These existing works did not consider
estimating a multivariate residual density and many of their sufficient conditions on the residual density are not easy to verify and some of them are even quite restrictive. Moreover, they also did not consider estimating the median function, which we believe more important for skewed response. Results on Bayesian consistency and convergence rates in multivariate response models are meagre with the exception of only completely nonparametric approaches to density estimation and mean regression (Shen and Ghosal (2011)). None of these methods deal with longitudinal data with a joint dependence structure induced by a copula. Also, these existing methods use either parametric or completely non-parametric mean function, whereas, for longitudinal response in biological and medical studies, it is more appropriate and interpretable to use a partial linear function for location, possibly with a non-parametric function for time and a parametric function for the effect of the main covariate of interest (e.g., HIV status for our data example). In addition, computations for the probit stick breaking mixtures of Pati and Dunson (2012) and the kernel stick breaking mixtures of Pelenis (2012) can be quite intensive. In this article, we obtain a computationally convenient semiparametric Bayes method that allows skewed multivariate residual density and leads to consistent estimation of the regression coefficients. In fact, we obtain our consistency results with only minor regularity conditions on the tail of the skewed and unimodal multivariate residual density.

This paper is organised as follows. In section 2, we describe the model and the estimators. Section 3 provides the theories and proofs for posterior consistency. Section 4 displays simulation study results about our model. Section 5 provides data analysis using this model on a real life dataset from the Pediatric Pulmonary and Cardiac Complications (P2C2) study of vertically transmitted HIV infection. Some remarks and discussions are in section 6.

### 1.2 Model Specifications

Let $y_i$ be the observed value of the longitudinally continuous response vector $Y_i$, $x_i$ be the covariate vector and $t_i$ be the vector of measurement time points for the subject $i = 1, \ldots, n$. We allow each subject to have different number of observations $m_i$ measured at different time points $t_{ij}$ for $j = 1, \ldots, m_i$. The regression model relating the responses variable $Y_{ij}$ with the predictors $x_{ij}$ at time $t_{ij}$ is given by

$$Y_{ij} = x_{ij}^T \beta + v(t_{ij}) + e_{ij},$$

where $x_{ij}$ is a $p \times 1$ covariate vector, $\beta$ is the corresponding regression parameter vector and $v(\cdot)$ is an unknown smooth function of time. Further, $e_i = (e_{i1}, \ldots, e_{im_i})$ is the error vector with median zero and possibly skewed common marginal density $f_e$. This ensures that the median of $y_{ij}$ is $m_{ij} = x_{ij}^T \beta + v(t_{ij})$.

There is a substantial literature proposing models for the nonparametric part $v(t)$ of a partial linear model Härdle et al. (2000) for a univariate response. Flexible representations include linear combinations of B-splines De Boor (1978), and penalized
regression splines Baladandayuthapani et al. (2005). However, none of these models aims to deal with correlated response variables. He et al. (2002) propose to model the nonparametric regression function via a normalised B-spline for longitudinal data. A spline is a sufficiently smooth piecewise-polynomial function, and a B-spline in particular has minimal support with respect to a given degree, smoothness and domain partition. In this paper, we use a B-spline to model the unknown smooth function \( v(t) \) in (1.1). Quantiles of observation time points \( t_{ij} \) are selected as the knots for, say, \( N \) b-spline functions \( b_1(t), \ldots, b_N(t) \). Then, the unknown function \( v(t) \) can be approximated by \( B(t)\xi, \) where \( B(t) = [b_1(t), \ldots, b_N(t)] \) and \( \xi \in \mathbb{R}^N \). Hence, our regression model becomes

\[
Y_{ij} = x_{ij}^T \beta + B(t_{ij})^T \xi + e_{ij},
\]

where \( e_{ij} \) is random variable with median zero and possibly skewed density function \( f_e(\cdot) \).

Kottas and Gelfand (2001) consider a special class of skewed densities for Bayesian analysis of univariate data, \( f(e|\theta, \lambda) = \frac{1}{\lambda} g_\theta(\frac{e}{\lambda}) I_{(-\infty, 0)}(e) + \lambda g_\theta(e) I_{(0, \infty)}(e), \) where \( I_{(\cdot)}(\cdot) \) is an indicator function, \( \lambda > 0 \) and \( g_\theta \) is a known unimodal density, Gaussian in particular, that is symmetric about 0. Motivated by Kottas and Gelfand (2001), our skewed error density \( f_e(\cdot) \) in (1.1) is assumed to be

\[
f_e(e|\lambda) = \begin{cases} 
    f_*(e)/2 & \text{if } e \geq 0, \\
    f_*(|e|/\lambda)/(2\lambda) & \text{if } e < 0,
\end{cases}
\]

where \( \lambda > 0 \) and \( f_*(u) \) is an unknown density function with support \((0, \infty)\). When \( \lambda \neq 1 \), any density of this family, is a skewed distribution with the type and amount of skewness depending on the value of \( \lambda \). When \( \lambda > 1 \), the resulting distribution is left skewed, when \( \lambda < 1 \), the resulting distribution is right skewed. We use the result by Feller (1966) that any decreasing density \( f \) with support \([0, \infty)\) can be expressed as

\[
f(x) = \int_{[0, \infty)} \frac{1}{y} I_{[0, y]}(x) G(dy) \tag{1.4}
\]

for some distribution \( G \) on \([0, \infty)\). As an immediate consequence of (1.4), when \( f_* \) in (1.3) is from a class of decreasing density function with mode at 0, then \( f_e(e) \) in (1.3) can be written as a mixture of uniform kernels:

\[
f_e(e|\lambda) = \int_0^\infty \left[ \frac{I_{[0, \theta]}(e)}{2\theta} + \frac{I_{(-\lambda\theta, 0)}(e)}{2\lambda\theta} \right] dG(\theta), \tag{1.5}
\]

where \( G(\theta) \) is an unknown cumulative density function with support \((0, \infty)\). We can rewrite the nonparametric class of \( F_e \) in (1.5) as

\[
F_e(e|\lambda) = \int_0^\infty \left[ I_{(\theta, \infty)}(e) + \frac{e + \theta}{2\theta} I_{[0, \theta]}(e) + \frac{e + \lambda\theta}{2\lambda\theta} I_{(-\lambda\theta, 0)}(e) \right] dG(\theta). \tag{1.6}
\]
To ensure the stationarity property of the marginal density \( f_e(\cdot) \) of the errors, which is a common feature in many longitudinal studies, we specify a multivariate distribution for error vector \( e_i \) that has common univariate marginal density \( f_e(\cdot) \) for each component while allowing flexible association over time within each subject. A copula model is a popular tool to specify a multivariate distribution with pre-specified marginal density \( f_e(\cdot) \) Nelsen (2006). The Gaussian copula Nelsen (2006); Pitt et al. (2006) is constructed using the probability integral transform of multivariate normal. We use a Gaussian copula for the vector of error \( e_i = (e_{i1}, \ldots, e_{im_i}) \) with a joint density

\[
\phi_{m_i}(\Phi_1^{-1}(F_e(e_{i1})), \ldots, \Phi_1^{-1}(F_e(e_{im_i}))) \prod_{j=1}^{m_i} \left| \frac{d\Phi_1^{-1}(F_e(e_{ij}))}{de_{ij}} \right|,
\]

where \( \phi_{m_i} \) is the probability density for a multivariate normal random vector with mean 0 and pre-specified covariance \( \Sigma \), \( \Phi_1^{-1} \) is the inverse of cumulative density for standard normal random variable.

Given observed data \( D = \{ y_i, x_i, t_i | i = 1, \ldots, n \} \), the likelihood is

\[
L(\beta, \xi, \lambda, \Sigma, f_*(\cdot)|D) \propto \prod_{i=1}^{n} \phi_{m_i}(e_i^*) \prod_{j=1}^{m_i} \left| \frac{de_{ij}^*}{de_{ij}} \right|,
\]

where \( \phi_{m_i}(e_i^*) \) is the mean 0 multivariate normal density with variance-covariance matrix \( \Sigma \), \( e_i^* = (\Phi_1^{-1}(F_e(e_{i1})), \ldots, \Phi_1^{-1}(F_e(e_{im_i}))) \) and \( e_{ij}^* = y_{ij} - x_{ij}^T \beta - B(t_{ij})^T \xi \). The Jacobian is a diagonal matrix with each element as

\[
\frac{de_{ij}^*}{de_{ij}} = \frac{d\Phi_1^{-1}(F_e(e_{ij}))}{de_{ij}} = \frac{1}{\phi_1(e_{ij}^*)} f_e(e_{ij}),
\]

for \( i = 1, \ldots, n \) and \( j = 1, \ldots, m_i \), where \( \phi_1(\cdot) \) is the probability density of standard normal distribution.

The maximum likelihood estimator of our parameters can be obtained via maximizing the log-likelihood function

\[
l(\beta, \xi, \lambda, \Sigma, f_*(\cdot)|D) = \sum_{i=1}^{n} \left[ \log(\phi_{m_i}(e_i^*)) + \sum_{j=1}^{m_i} \log(f_e(e_{ij}^*)) - \sum_{j=1}^{m_i} \log(\phi_1(e_{ij}^*)) \right].
\]

The secant method with numerical differentiation for the gradient can be applied to get the MLE with parametric \( f_*(\cdot|\gamma) \), where \( \gamma \) is the unknown parameter of \( f_*(\cdot) \) in equation (1.3).

For a parametric model with, say, exponential density for \( f_*(\cdot) \), the parametric likelihood \( L(\beta, \xi, \lambda, \Sigma, f_*(\cdot)|D) \) is the same as (1.7) with \( f_e(e) \) replaced by

\[
f_e(e) = \begin{cases} 
\gamma \exp\{-\gamma e\}/2 & \text{if } e \geq 0, \\
\gamma \exp\{\gamma e/\lambda\}/(2\lambda) & \text{if } e < 0.
\end{cases}
\]
The posterior density is 
\[ p(\beta, \xi, \lambda, \Sigma, \gamma | \mathbf{D}) \propto L(\beta, \xi, \lambda, \Sigma, \gamma | \mathbf{D}) \pi(\beta, \xi, \lambda, \Sigma, \gamma), \]
where \( \pi(\beta, \xi, \lambda, \Sigma, \gamma) \) is the joint prior density based on the available prior information. The parameters \( \beta, \xi, \lambda, \Sigma, \gamma \) are supposed to have mutually independent prior densities. The resulting full Bayesian model has the hierarchical structure with conditional density \( [ \mathbf{Y}_i | \beta, \xi, \lambda, \Sigma, \gamma ] \) of (1.2) and \( (\beta, \xi, \lambda, \Sigma, \gamma) \sim \pi(\beta, \xi, \lambda, \Sigma, \gamma) \). Markov Chain Monte Carlo (MCMC) samples from this joint posterior density can be used to implement a parametric Bayesian analysis of this hierarchical model.

However, the parametric assumption about \( f_e(\cdot) \) is restrictive for some studies in practice. For the semiparametric model of (1.2), we use the class of skewed distributions given in (1.5). The semiparametric likelihood of this model is the same as (1.7) with parametric \( f_e(e) \) replaced by the nonparametric \( f_e(\cdot) \) of (1.5) with unknown \( G \). For a semiparametric Bayesian analysis, the posterior density is
\[ p(\beta, \xi, \lambda, \Sigma, G | \mathbf{D}) \propto L(\beta, \xi, \lambda, \Sigma, G | \mathbf{D}) \pi_1(\beta, \xi, \lambda, \Sigma) \pi_2(G), \]  
where the prior process for \( F_e(\cdot | \lambda) \) is based on a prior process \( \pi_2(G) \) of an unknown mixing distribution \( G \) of (1.5). We use the Dirichlet process Ferguson (1973, 1974) \( G \sim DP(\alpha, G_0) \) as the prior \( \pi_2(G) \) for the unknown mixing distribution \( G(\theta) \) of (1.5), where \( G_0 \) is the prior mean ("prior guess") of unknown \( G \) and \( \alpha \) is the pre-specified precision around "prior guess". One big practical advantage of our semiparametric Bayes analysis is that the \( DP(\alpha, G_0) \) prior process on \( G \) of (1.5) can be specified using the "prior guess" of marginal error \( F_e \), because the \( G_0 \) corresponds to an unique "prior guess" \( F_{e,\text{prior}}(\cdot | G_0, \lambda = 1) = E[F_e(\cdot | \lambda = 1) | G_0] \) of \( F_e(\cdot | \lambda = 1) \) via \( G_0'(u) = -uF_{e,\text{prior}}'(u) \) for \( u > 0 \) (Khintchine, 1938) proof follows from the proof of a similar result by. For example, to obtain an approximate double exponential prior mean density for \( f_e(\cdot | \lambda = 1) \), we need to choose \( G_0 \) as a gamma distribution. When \( G_0 \) is inverse gamma, the \( F_{e,\text{prior}}(\cdot | G_0, \lambda = 1) \) turns to be a heavy tailed distribution. One of the major advantages of our method is the ease of determination of the prior \( \pi_2(G) \) based on "prior guess" for \( F_e(\cdot | \lambda = 1) \) and assuming a prior that \( \pi_2(F_e) = \pi_2(G, \lambda) \approx \pi_3(\lambda) \pi_4(G | \lambda = 1) \). The resulting full Bayesian model has a hierarchical structure with (i) \( [ \mathbf{Y}_i | \beta, \xi, \lambda, \Sigma, \theta_i ] \) having i.i.d. distribution of (1.2), (ii) \( \theta_i | G \sim G \) for \( j = 1, \ldots, m_i \) and \( G \sim DP(\alpha, G_0) \); (iii) \( (\beta, \xi, \lambda, \Sigma) \sim \pi(\beta, \xi, \lambda, \Sigma) \).

The constructive definition of the DP mixture prior process for \( F_e \) Sethuraman (1994); Ishwaran and James (2001) can be represented as
\[ F_e(e) = \sum_{k=1}^{\infty} p_k \left[ I(\theta_k, \infty)(e) + \frac{\theta_k + e}{2\theta_k} I(0, \theta_k)(e) + \frac{e + \lambda \theta_k}{2\lambda \theta_k} I(-\lambda \theta_k, 0)(e) \right], \]  
where \( \theta_k \overset{i.i.d.}{\sim} G_0, p_1 = V_1, p_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1})V_k, k \geq 2, \) and \( V_k \overset{i.i.d.}{\sim} Beta(1, \alpha) \) random variables. In this article, a finite approximation of \( f_e \) and \( F_e \) of the form described in (1.11) are implemented using MCMC.
1.3 Properties of Prior and Posterior Consistency

Recall that our model for the observations \( Y_{ij} \) is

\[
Y_{ij} = x_{ij}^T \beta + v(t_{ij}) + e_{ij},
\]

where \( i = 1, \ldots, n \) for subject, \( j = 1, \ldots, m_i \) for different observations within subject \( i \). For technical simplicity, it is reasonable to assume that \( m_i \)'s are balanced with \( m_i \equiv m \) and \( \beta \) is univariate and henceforth denoted by \( \beta \). Below, we study the support of this prior and provide conditions under which we obtain consistent estimates of the coefficients and the unknown function of time.

1.3.1 Large support of the prior for the multivariate residual density

In this section, we will formalize the notions of prior support and demonstrate that a large class of residual densities lies in the support of our proposed prior. For studying posterior consistency, we will assume that the true data generating distribution lies in this subset - so it is crucial to ensure that the subsets are large enough to cover a wide range of true distributions.

To that end, we first define \( F_{au} \) be the set of all univariate asymmetric unimodal densities with median zero with an element in \( F_{au} \) satisfying \( 1/2f_\ast(e)I(e>0) + 1/(2\lambda_\ast)f_\ast(|e|/\lambda_\ast)I(e < 0) \) for some decreasing density \( f_\ast \) on \([0, \infty)\) and for some \( \lambda_\ast > 0 \). Let \( F_\rho \) be the set of densities supported on \((0, 1)\). Define \( C_G((F_{au})^m \times F_\rho) \) to be the set of all multivariate densities arising from \( m \) univariate median zero residual densities based on Gaussian copula, i.e., any \( f_0 \in C_G((F_{au})^m \times F_\rho) \) can be expressed as

\[
f_0(e) = \tilde{\phi}(\rho | \Sigma_0) \prod_{j=1}^m f_{0j}(e_j),
\]

where \( e = (e_1, \ldots, e_m)^T, \Sigma_0 \) is a \( m \times m \) positive definite correlation matrix with \( \rho_0 \in F_\rho, f_{0j} \in F_{au} \) and \( \phi(\rho | \Sigma_0) = \frac{1}{\sqrt{|\Sigma|}} \exp\left\{-\rho^*/(\Sigma_0^{-1} - I)\rho^*/2\right\} \), where \( \rho^* = \Phi^{-1}(F_{0j}(e_j)) \) and \( F_{0j} \) is the cdf of \( f_{0j} \).

Assuming the prior on \( F_{au} \) and \( F_\rho \) to be \( \Pi_{au} \) and \( \Pi_\rho \) respectively, we denote the prior for \( f \in C_G((F_{au})^m \times F_\rho) \) by \( \Pi_{C_G} = (\Pi_{au})^m \otimes \Pi_\rho \). Recall that \( \Pi_{au} \) is constructed by letting

\[
f(e | G, \lambda) = \int_{\mathbb{R}^+} \frac{1_{(-\lambda\theta, 0)}(e)}{2\lambda \theta} + \frac{1_{(0, \theta)}(e)}{2\theta} G(d\theta), \ G \sim DP(\alpha, G_0), \ \lambda \sim \Pi_\lambda(1.12)
\]

Next, we will define various neighbourhoods of \( f_0 \) to demonstrate the support of the prior for the residual density. The Kullback-Leibler neighbourhood, denoted by \( KL(f_0; f) = \int_{\mathbb{R}^n} f_0 \log \frac{f_0}{f} \) is a non-symmetric measure of distance between two
densities and plays a crucial role guaranteeing that the prior is flexible enough in approximating the true distribution. Recall that the Kullback-Leibler support of a prior $\Pi$ on a density space $\mathcal{F}$, denoted by $KL(\Pi)$, is defined by the subset of $\mathcal{F}$ satisfying

$$f_0 : \Pi(f : KL(f_0; f) < \epsilon) > 0 \text{ for all } \epsilon > 0 \right\}.$$

Denote by $wk(\Pi)$ the weak support of the prior $\Pi$, i.e, the support of the prior $\Pi$ under weak topology. Define $\mathcal{F}_{KL} = \{f_0 \in C_G((\mathcal{F}_{au})^m \times \mathcal{F}_{e})\}, \int (\sum_{j=1}^{m} |\log f_{0j}|)^2m \Pi_{j=1}^{m} f_{0j} < \infty\}$. In the following Lemma 1, we characterize the weak and the Kullback-Leibler support of $\Pi_G$.

Lemma 1. (1) $wk(\Pi_G) = C_G((\mathcal{F}_{au})^m \times \mathcal{F}_e)$.

(2) $KL(\Pi_G) \supset \mathcal{F}_{KL}$ if $\text{supp}(G_0) = \mathbb{R}^+, G_0$ is absolutely continuous with respect to the Lebesgue measure, $\text{supp}(\Pi_\Lambda)$ is the set of all distribution on $\mathbb{R}^+$ and $\text{supp}(\Pi_\theta) = \mathcal{F}_\theta$.

Remark 1. As the KL distance is a stronger notion of divergence compared to the weak topology, the tails of the residual density should decay at the rate specified by $\int (\sum_{j=1}^{m} |\log f_{0j}|)^2m \Pi_{j=1}^{m} f_{0j} < \infty\}$ so that it can be approximated by the prior realizations in KL distance.

Proof: We only provide the details for 2) as the proof for 1) is straightforward. Given a density $f_0 \in \mathcal{F}_{KL}$, the idea is to construct a sequence of functions $f_M \in wk(\Pi_G), M \geq 1$ such that $KL(f_0, f_M) \to 0$ as $M \to \infty$. Define $f_M = \phi_M \Pi_{j=1}^{m} f_{jM}$ where $KL(f_{0j}, f_{jM}) \to 0$ as $M \to \infty$ for each $j = 1, \ldots, m$ and

$$\phi_M = \frac{1}{\sqrt{|\Sigma_M|}} \exp\{-\Phi_1^{-1}(F_M(e))'(\Sigma_M^{-1} - I)\Phi_1^{-1}(F_M(e))\},$$

where $\Phi_1^{-1}(F_M(e)) = (\Phi_1^{-1}(F_{1M}(e_1)), \ldots, \Phi_1^{-1}(F_{mM}(e_m)))$, $F_{jM}$ is the cdf of $f_{jM}$ for $j = 1, \ldots, m$ and $\Sigma_M$ is an uniformly correlation matrix with $\rho_M \to \rho_0$. We construct $f_{jM}$ below. Let $g_{0j}(e) = 2f_{0j}^+(e), g_{0j}(-e/\lambda_0)/\lambda_0 = 2f_{0j}^-(e)$. Clearly $g_{0j}$ is a continuous decreasing density on $\mathbb{R}^+$ and $f_{0j}(e) = 0.5(g_{0j}(e) + g_{0j}(-e/\lambda_0)/\lambda_0)$. As in Wu and Ghosal (2008), we let $e_1 > 0$ and $e_2 > 0$ such that $g_{0j}(e_1) = a$ and $g_{0j}(e_2) = b$, where $0 < b < 1$ and $b < a < g_{0j}(0)$. For given $M$, let $M_1$ and $M_2$ be such that $\frac{M_1}{M} \leq e_1 \leq \frac{M_1+1}{M}$ and $\frac{M_2}{M} \leq e_2 \leq \frac{M_2+1}{M}$. Set

$$w_{ji}^* = \begin{cases} \frac{i}{M} \{g_{0j}\left(\frac{i}{M}\right) - g_{0j}\left(\frac{i+1}{M}\right)\}, & 1 \leq i < M_1, \\ \frac{M_1}{M} \{g_{0j}\left(\frac{M_1}{M}\right) - a\}, & i = M_1, \\ \frac{M_1+1}{M} \{a - g_{0j}\left(\frac{M_1+1}{M}\right)\}, & i = M_1 + 1, \\ \frac{i}{M} \{g_{0j}\left(\frac{i+1}{M}\right) - g_{0j}\left(\frac{i}{M}\right)\}, & M_1 + 1 < i \leq M_2, \\ \frac{M_2}{M} \{g_{0j}\left(\frac{M_2}{M}\right) - g_{0j}\left(\frac{i}{M}\right)\}, & i \geq M_2 + 1. \end{cases}$$

We define $f_{jM}^*(e) = \sum_{i}^\infty w_{ji}^* K(e; \frac{i}{M})$, where $K(e; \theta) = \theta^{-1}1_{(0 \leq e \leq \theta)}$. By the continuity of $g_0, f_{jM}^*$ converges to $g_0$ pointwise. Note that $f_{jM}^*$ is not a p.d.f. Let
\[
\begin{align*}
\text{w}_{ji} &= w_{ji}^* \frac{1-M_{i-1}w_{ji}-M_jw_{ji}}{M_jw_{ji}}, \quad \text{for } M_1 \leq i \leq M_2 \text{ and } w_{ji} = w_{ji}^* \text{ for all other } i.'s.
\end{align*}
\]

Then \( w_{ji}'s \) are positive and \( \sum_{i=1}^{\infty} w_{ji} = 1 \). Let \( f_{JM}^+(e) = \sum_{i=1}^{\infty} w_{ji}K(e; i_M) \). Clearly, \( f_{JM}^+(e) \rightarrow f_{JM}(e) \) as \( M \rightarrow \infty \) by the definition of Riemann integral. Thus \( f_{JM}^+(e) \) converges to \( g_0(e) = 2f_0(e)I(e>0) \) pointwise. We also have

\[
|\log f_{JM}^+(e)| \leq \begin{cases} 
\max(|\log 2 + |\log g_0|, |\log a - \log 2|), & 0 < e \leq M_1 + 1, \\
\max(|\log a, (|\log g_0(e_2 + 1)|)), & M_1 + 1 < e \leq M_2 + 1, \\
|\log g_0|, & e > M_2 + 1.
\end{cases}
\]

If \( f_0 \in F_{KL} \), we have \( \int f_0|\log f_0| < \infty, \forall j \). Hence \( \int_{\mathbb{R}^+} f_0^+|\log f_0| < \infty, \) and \( \int_{\mathbb{R}^+} f_0^+|\log f_0| < \infty \). Since \( |\log g_0(e)| \) is \( g_0 \)-integrable, using DCT, we have the \( \int g_0|\log g_0| f_{JM}^* \rightarrow 0 \) as \( M \rightarrow \infty \). Defining \( f_{JM}(e) = \frac{1}{2}(\phi_0(e) + \phi_0(e)) - \frac{1}{2\lambda_M} f_{JM}^* \) with \( \lambda_M \rightarrow \lambda_0 \), \( |\log f_{JM}(e)| \) is bounded above by an \( f_0 \)-integrable, and hence the \( \int f_0|\log f_0| \rightarrow 0 \) as \( M \rightarrow \infty \).

Below, we show that \( f_M \rightarrow f_0 \) pointwise and construct an \( f_0 \)-integrable upper bound of \( |\log f_M| \). Clearly, since \( f_{JM} \rightarrow f_0j \) pointwise for \( j = 1, \ldots, m, \) \( \sup_{e \in \mathbb{R}} |F_{JM}(e) - F_{0M}(e)| \rightarrow 0 \) as \( M \rightarrow \infty \) by Scheffe's theorem. As \( \Phi^{-1} \) and \( \phi \) are both continuous in their arguments, \( f_M \rightarrow f_0 \) pointwise as \( M \rightarrow \infty \). Observe that

\[
|\log f_M| \leq |\log \tilde{\phi}_M| + \sum_{j=1}^{m} |\log f_{JM}|.
\]

Hence \( \sum_{j=1}^{m} |\log f_{JM}| \) is bound above by an \( f_0 \)-integrable function. It remains to show the same for \( |\log \tilde{\phi}_M| \). We have

\[
\log \tilde{\phi}_M = -\Phi^{-1}(F_M(e))'(\Sigma^{-1} - I)\Phi^{-1}(F_M(e)) - \log |\Sigma_M|^2 / 2,
\]

\[
= -\Phi^{-1}(F_M(e))'(\Sigma^{-1} - I)\Phi^{-1}(F_M(e)) + c_1,
\]

\[
= -\Phi^{-1}(F_M(e))'(\Sigma^{-1} - I)\Phi^{-1}(F_M(e)) + c_1.
\]

For \( j = 1, \ldots, m, \) we have \( e_{JM}^* = \Phi^{-1}(F_{JM}(e)), \) \( e_{0j} = \Phi^{-1}(F_{0j}(e)) \). Using Taylor series expansion of \( e_{JM}^* \), we divide for some \( \xi \in [0, 1],
\]

\[
e_{JM} = e_{0j} + \frac{F_{JM}(e) - F_{0j}(e)}{\phi_1(F_{0j}(e))} + \frac{\phi_1'(\xi)(F_{JM}(e) - F_{0j}(e))^2}{\phi_1^2(\xi)}.
\]

Hence \( |e_{JM} - e_{0j}| \leq C|F_{JM}(e) - F_{0j}(e)| \rightarrow 0 \) uniformly in \( e \) making the first term in (1.14) uniformly small in \( e \) as \( M \rightarrow \infty \). Since \( \rho_M \rightarrow \rho_0, \) the \( m \) eigenvalues of \( \Sigma^{-1} \) will converge to the \( m \) eigenvalue of \( \Sigma_0^{-1} \), making \( e_0'(\Sigma_0^{-1} - \Sigma^{-1})e_0^* / 2 \leq \max |\lambda_0 - \lambda_M|(e_0'(\Sigma_0^{-1} - \Sigma^{-1})e_0^* \rightarrow 0 \) as \( M \rightarrow \infty \). Thus,

\[
|\log \tilde{\phi}_M| \leq c_1 + c_2|\log \tilde{\phi}_0|.
\]
which is $f_0$-integrable as $|\log \tilde{\phi}_0|$ is $f_0$-integrable by change of variable. From (1.13) and (1.15), $|\log f_M|$ is bounded by an $f_0$ integrable function. By DCT, we can conclude that $\mathcal{F}_{KL} \cap \{f_0 \in \mathcal{C}_G((\mathcal{F}_{au})^m \times \mathcal{F}_f), \int f_0 |\log f_0| < \infty\} \subset KL(\Pi_{CG})$.

Now, we show $\mathcal{F}_{KL} \subset \{f_0 \in \mathcal{C}_G((\mathcal{F}_{au})^m \times \mathcal{F}_f), \int f_0 |\log f_0| < \infty\} \subset KL(\Pi_f)$. Using Holder’s inequality for any $\tilde{\alpha} > 1$, we have

$$
\int f_0 |\log f_0| = |\int \tilde{\phi}(\Sigma) \prod_{j=1}^{m} f_{0j} |\log \tilde{\phi}(\Sigma) \prod_{j=1}^{m} f_{0j}|,
$$

(1.16)

$$
= \int \tilde{\phi}(\Sigma) \left( \prod_{j=1}^{m} f_{0j} \right)^{1/\tilde{\alpha}} |\log \tilde{\phi}(\Sigma) \prod_{j=1}^{m} f_{0j}| \left( \prod_{j=1}^{m} f_{0j} \right)^{1-1/\tilde{\alpha}},
$$

(1.17)

$$
\leq \left[ \int \tilde{\phi}^{\tilde{\alpha}}(\Sigma) \prod_{j=1}^{m} f_{0j} \right]^{1/\tilde{\alpha}} \left[ \int |\log \tilde{\phi}(\Sigma) \prod_{j=1}^{m} f_{0j}| \prod_{j=1}^{m} f_{0j} \right]^{1-1/\tilde{\alpha}},
$$

(1.18)

$$
= \left[ \int \frac{1}{|\Sigma|^{\tilde{\alpha}/2}} \exp\{-e^{*}(\tilde{\alpha} \Sigma^{-1} - (\tilde{\alpha} - 1)I) e^{*}\} d e^{*}\right]^{1/\tilde{\alpha}} \times \left[ \int |\log \tilde{\phi}(\Sigma) \prod_{j=1}^{m} f_{0j}| \prod_{j=1}^{m} f_{0j} \right]^{1-1/\tilde{\alpha}}.
$$

(1.19)

We will choose $\tilde{\alpha}$ to make the first term in (1.19) is finite. For an $m \times m$ positive definite matrix with compound symmetry correlation structure (with correlation element $\rho$), the eigen values are $1 - \rho, \ldots, 1 - \rho$ and $1 + (m - 1) \rho$ with the associated eigenvectors $(1, \frac{-1}{m-1}, \ldots)^T$, $(\frac{1}{m-1}, \ldots, 1)^T$ and $1_m$. Hence $\Sigma^{-1}$ has eigenvalues $\frac{1}{1-\rho}$ and $\frac{1}{1+(m-1)\rho}$. Choosing $\tilde{\alpha} = 1 + \frac{1}{m-1}$, $\tilde{\alpha} \Sigma^{-1} - (\tilde{\alpha} - 1)I$ is positive definite and hence we can make the first term finite for any $\rho \in (0, 1)$.

The second part in (1.19) can be bounded above by Minkowski’s inequality as follows.

$$
\left\{ \int |\log \tilde{\phi}(\Sigma) \prod_{j=1}^{m} f_j|^{m} \prod_{j=1}^{m} f_j \right\}^{1/m} \leq \left\{ \int |\log \tilde{\phi}(\Sigma)|^{m} \prod_{j=1}^{m} f_j \right\}^{1/m} + \left\{ \int \log \prod_{j=1}^{m} f_j |\prod_{j=1}^{m} f_j|^{1/m} \right\}.
$$

(1.20)
The first part on the right of (1.20) can again be upper bounded as follows.

\[
\int |\log \tilde{\phi}(\Sigma)|^m \prod_{j=1}^m f_{0j} \leq C_1 + C_2 \int |(\mathbf{e}^*)'(\Sigma^{-1} - I)(\mathbf{e}^*)|^m \prod_{j=1}^m f_{0j}(e_j)
\]

\[
= C_1 + C_2 \int |(\mathbf{e}^*)'(\Sigma^{-1} - I)(\mathbf{e}^*)|^m \prod_{j=1}^m \phi_1(e_j^*),
\]

\[
\leq C_1 + C_3 \int |(\mathbf{e}^*)'(\Sigma^{-1})(\mathbf{e}^*)|^m \prod_{j=1}^m \phi_1(e_j^*)
\]

\[
+ C_4 \int |(\mathbf{e}^*)'(\mathbf{e}^*)|^m \prod_{j=1}^m \phi_1(e_j^*),
\]

\[
\leq C_1 + C_5 \int |(\mathbf{e}^*)'(\mathbf{e}^*)|^m \prod_{j=1}^m \phi_1(e_j^*)
\]

by change of variable. This is always finite as the multivariate Gaussian has moments of all orders. Thus, we have

\[
\left\{ f_0 \in G ( (\mathcal{F}_{aa})^m \times \mathcal{F}_\theta), \int (\sum_{j=1}^m |\log f_{0j}|)^m \prod_{j=1}^m f_{0j} < \infty \right\}
\]

\[
\subset \{ f_0 \in \mathcal{F}_{KL}, \int f_0|\log f_0| < \infty \}.
\]

**Remark 2.** The uniform correlation matrix can be extended to any positive definite correlation matrix. For positive correlation matrix, the sum of eigenvalues \(\lambda_1 + \ldots + \lambda_m = m\), \(\lambda_j > 0, j = 1, \ldots, m\). Thus, we have \(\lambda_j < m\) and \(\bar{a}\Sigma^{-1} - (\bar{a} - 1)I\) is positive definite.

### 1.3.2 Spline approximation property

In this section we will show that a B-splines prior is sufficiently flexible to approximate any unknown continuous function of time for suitably chosen prior distributions for the number of knots. Assume that \(t_{ij} \in [A, B]\) for all \((i, j)\). For a given natural number \(K\), partition the interval \([A, B]\) into \(K\) subintervals using knot points \(t_1 = \cdots = t_{q+1} = A < t_{q+2} < t_{q+3} < \cdots < t_{q+K} < t_{q+K+1} = \cdots = t_{2q+K+1} = B\). Let \(J = (q + K)\) and \(\mathcal{B}_{q,J} = \{ b_{q,1}, b_{q,2}, \ldots, b_{q,J} \}\) denote B-spline bases of degree \(q\). For any \(x \in [A, B]\) the B-splines satisfy (i) \(b_{q,j}(x) \geq 0 \forall j\), (ii) \(\sum_{j=1}^J b_{q,j}(x) = 1\), (iii) \(b_{q,j}\) is supported inside the interval \([t_j, t_{j+q+1}]\), and (iv) for any \(x \in (t_j, t_{j+1})\), \((q + 1)\) B-splines \(b_{q,j-q}(x), b_{q,j-q+1}(x), \ldots, b_{q,j}(x)\) are non-zero. Define \(S_{q,J} = \{ v_s : v_s = \mathcal{B}_{q,J} \xi = \sum_{j=1}^J b_{q,j} \xi_j \text{ for some } \xi \in \mathbb{R}^J \}\).

Let \(\mathcal{C}[A, B]\) denote the set of continuous functions from \([A, B]\) to \(\mathbb{R}\). Let \(\mathcal{C}^\kappa[A, B] \subset \mathcal{C}[A, B]\) denote the set of functions that are \(\kappa_0\) times continuously differentiable and,
for any $v \in C^\kappa[A, B]$, $\|v\|_\kappa < \infty$, where $\kappa_0$ is largest integer less than $\kappa$ and the seminorm is defined by $\|v\|_\kappa = \sup_{x, x' \in [A, B], x \neq x'} \{|v|^{(\kappa_0)}(x) - v^{(\kappa_0)}(x')|/|x - x'|^{\kappa - \kappa_0}\}$. Let $\|v\|_\infty = \sup_{x \in [A, B]} |v(x)|$, and let $\|\cdot\|_2$ denote the Euclidean norm.

Given $q$, let $\Pi_q$ denote a prior on $\mathbb{N}_q = \{q + 1, q + 2, \ldots\}$ such that $\Pi_q(J) > 0 \forall J \in \mathbb{N}_q$. Given $J \sim \Pi_q$, also let $\Pi_{q, J}$ be a prior on $\mathbb{R}^J$ such that $\Pi_{q, J}(N_\delta(\xi_0)) > 0$ for any $\delta > 0$ and any $\xi_0 \in \mathbb{R}^J$, where $N_\delta(\xi_0) = \{\xi : \xi \in \mathbb{R}^J, ||\xi - \xi_0||_2 < \delta\}$. Let $\Pi_v = \Pi_q \times \Pi_{q, J}$ denote the induced prior on $S_q = \bigcup_{q=0}^\infty S_{q, J}$.

Lemma 2. For any $v_0 \in C[A, B]$ (or $C^\kappa[A, B]$), $\Pi_v(\{v \in S_q : ||v - v_0||_\infty < \epsilon\}) > 0 \forall \epsilon > 0$.

Define $\psi(v, h) = \sup_{x, x' \in [A, B], |x-x'| \leq h} |v(x) - v(x')|$. Define $\tau_j = (t_j + t_{j+q+1})/2$, the midpoint of the support of $b_{q, j}$ for $j = 1, 2, \ldots, J$. Also let $\Delta_j = (t_{j+1} - t_j)$ for $j = 1, 2, \ldots, (J-1)$. The $q^{th}$ order spline approximation $\hat{v}(\cdot, \tau) \in S_{q, J}$ to a function $v \in C[A, B]$ is defined as $\hat{v}(\cdot, \tau) = \sum_{j=1}^J v(\tau_j)b_{q, j}(\cdot)$. Since B-splines add up to one, the transformation reproduces constants, that is, $\hat{v}_s = v$, in case $v$ is a constant, $v(x) = c > 0$ for all $x \in [A, B]$. In general, for $v \in C[A, B]$, for any $x_0 \in [t_j, t_{j+1}]$,

$$\left|\hat{v}_s(x_0, \tau) - v(x_0)\right| \leq \sum_{j=1}^J \left|v(\tau_j)b_{q, j}(x_0) - v(x_0)b_{q, j}(x_0)\right| \leq \sup_{(j-q) \leq k \leq j} \left|v(\tau_k) - v(x_0)\right| \leq \sup\{|v(x) - v(x')| : x, x' \in [t_j, t_{j+1}]\} \leq \psi(v, (q + 1)\Delta_{\max}/2) \leq [(q + 1)/2] \psi(v, \Delta_{\max}),$$

where $\Delta_{\max} = \max_{1 \leq j \leq (J-1)} \Delta_j$ and $[\kappa] = \min\{n : n \in \mathbb{N}, n \geq \kappa\}$. The second last inequality follows from the facts that $(t_{j+1} - \tau_{j-q}) = (t_{j+1} + (t_{j+1} + t_j - t_{j-q})/2) = (t_{j+1} - t_{j-q})/2 \leq ((q + 1)/2) \Delta_{\max}$ and $(\tau_j - t_j) = ((t_{j+1} + t_j + t_{j+1} + t_{j+1} + t_{j+1} - t_j)/2 - t_j) = (t_{j+q+1} - t_j)/2 \leq ((q + 1)/2) \Delta_{\max}$. The last step uses the monotonicity and subadditivity of $\psi(v, \cdot)$. We conclude

$$\inf_{v_s \in S_{q, J}} ||v - v_s||_\infty \leq ||v - \hat{v}(\cdot, \tau)||_\infty \leq c(q) \psi(v, \Delta_{\max}) \to 0 \text{ as } \Delta_{\max} \to 0.$$
application of these three facts gives

\[ \inf_{v_s \in S_{q,j}} \|v - v_s\|_\infty = \inf_{v_s', v_s'' \in S_{q,j}} \|v - v_s - v_s'\|_\infty \leq c(q) \inf_{v_s \in S_{q,j}} \psi(v - v_s, \Delta_{\max}) \]

\[ \leq c(q) \Delta_{\max} \inf_{v_s \in S_{q,j}} \|v^{(1)} - v_s^{(1)}\|_\infty \]

\[ = c(q) \Delta_{\max} \inf_{v_s \in S_{q,j}} \|v^{(1)} - v_s\|_\infty \]

\[ \leq c(q) c(q-1) \Delta_{\max} \psi(v^{(1)}, \Delta_{\max}) \leq \ldots \]

\[ \leq c(q) c(q-1) \ldots c(q - \kappa_0 + 1) \Delta_{\max}^{\kappa_0-1} \psi(v^{(\kappa_0-1)}, \Delta_{\max}) \]

\[ \leq c(q) c(q-1) \ldots c(q - \kappa_0 + 1) \Delta_{\max}^{\kappa_0} \|v^{(\kappa_0)}\|_\infty. \]

For any two functions \(g_1\) and \(g_2\), \(\sup |g_1| \leq \sup |g_1| \sup |g_2|\). Taking \(g_1(x, x') = \{v^{(\kappa_0)}(x) - v^{(\kappa_0)}(x')\}/(x - x')^{(\kappa - \kappa_0)}\) and \(g_2(x, x') = (x - x')^{(\kappa - \kappa_0)}\), we have \(||v^{(\kappa_0)}\|_\infty \leq ||v||_\kappa (B - A)^{(-\kappa_0)}\). Therefore, when the knot points \(\{t_{q+j}\}_{j=0}^K\) are equidistant

\[ \inf_{v_s \in S_{q,j}} \|v - v_s\|_\infty \leq c(q, \kappa_0) \frac{(B - A)^{\kappa_0}}{K^{\kappa_0}} \|v^{(\kappa_0)}\|_\infty \]

\[ \leq c(q, \kappa_0) \frac{(B - A)^{\kappa}}{K^{\kappa}} \|v\|_\kappa = c(q, \kappa) K^{-\kappa} \|v\|_\kappa. \]

Given any \(v_0 \in C_+[A, B]\) (or \(C^\kappa[A, B]\)) and \(\epsilon > 0\), find \(J \in \mathbb{N}_q\) and \(\xi_0 \in \mathbb{R}^J\) such that \(||v_0 - B_{q,j}\xi_0||_\infty = \inf_{v_s \in S_{q,j}} \|v - v_s\|_\infty < \epsilon/2\). Next consider a neighbourhood \(N_\delta(\xi_0)\) such that for any \(\xi \in N_\delta(\xi_0)\), we have \(||B_{q,j}\xi - B_{q,j}\xi_0||_\infty < \epsilon/2\). Then for any \(\xi \in N_\delta(\xi_0)\), we have \(||B_{q,j}\xi - v_0||_\infty \leq ||B_{q,j}\xi - B_{q,j}\xi_0||_\infty + ||B_{q,j}\xi_0 - v_0||_\infty < \epsilon\). Then \(\Pi_\epsilon(||v - v_0||_\infty < \epsilon) \geq \Pi_q(J) \Pi_{\xi\mid J}\{N_\delta(\xi_0)\} > 0\). Hence the proof of Lemma 2.

Remark 3. When \(v_0\) is a combination of B-splines, the prior for \(J\) induced through \(\Pi_q\) is not required. For any \(\epsilon > 0\), we have \(\Pi_v(\{v \in S_q : ||v - v_0||_\infty < \epsilon\}) > 0\) for large enough \(J\).

### 1.3.3 Posterior consistency

In this section, we will provide conditions that guarantee that as we get more and more samples from the true data generating distribution, the posterior distribution of the unknown quantities concentrate around arbitrary small pre-defined neighbourhoods around the truth. The choice of neighbourhoods is decided from practical considerations. In the partial linear model framework, the primary interest is in estimating the fixed effects \(\beta\) and the residual density can be considered as a nuisance. As we are modeling the unknown function of time non-parametrically, we certainly cannot hope to recover the true function apart from the function values at the observed time points. A stronger consistency result ensuring recovery of the function at unobserved time points will require additional assumptions on the design on the sampled time points (see Amewou-Atisso et al. (2003)). Such assumptions are often unrealistic and are hence not considered here.
Define \( f_{0i} = f_0(y - v_0(t_i) - \beta_0 x_i | \lambda_0) \), \( \eta_0 = (v_0, \beta_0) \), \( \eta = (v, \beta) \), \( f_{\eta} = f(y - v(t_i) - \beta x_i | \lambda) \), where \( t_i \) and \( x_i \) are the vector measurement of time and covariate for subject \( i \). Also, for any two multivariate densities \( f \) and \( \tilde{g} \) with dimension \( m \), we let \( K(f, \tilde{g}) = \int f(y) \log \left\{ \frac{f(y)}{g(y)} \right\} dy \), \( V(f, \tilde{g}) = \int f(y) \log \left\{ \frac{f(y)}{g(y)} \right\}^2 dy \), and \( K_i(\eta, \eta) = K(f_{0i}, f_{\eta i}) \), \( V_i(\eta, \eta) = V(f_{0i}, f_{\eta i}) \). We consider a strong neighbourhood around the true fixed effect \( \beta_0 \), an \( L_2 \) type norm based on the empirical distribution for the unknown function of time and a weak neighbourhood for the residual density.

Consider \( U = U \times \{ v \in S_q : \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} |v(t_{ij}) - v_0(t_{ij})| < \epsilon \} \times \{ \beta : \left| \beta - \beta_0 \right| < \epsilon \} \), where \( U \) is a weak neighbourhood of \( f_0 \).

**Theorem 1.** Assume that \((f_0, v_0, \beta_0) \in \mathcal{F}_{KL} \times C^\alpha[A, B] \times \mathbb{R} \). Consider the prior \( \Pi = (\Pi_G \times \Pi_v \times \Pi_\beta) \) on \( \mathcal{F}_{KL} \times C^\alpha[A, B] \times \mathbb{R} \). Assume that the prior for \( J \) satisfies \( P(J \geq J_n) \leq e^{-nCJ} \) with \( J_n = o(n) \) for some constant \( C_J > 0 \). Then

\[
\Pi\{ (f, \eta) \in U^c | (Y_1, x_1), \ldots, (Y_n, x_n) \} \to 0 \text{ a.s.}[P_{f_0, v_0}].
\]

**Proof.** The conclusion of Theorem 1 holds following Pati and Dunson (2012) provided there exists test functions \( \{ \Phi_n(\{y_i, t_i, x_i\}_{i=1}^{n}) \}_{n=1}^{\infty} \), sets \( \Theta_n = U \times \Theta_v \times \Theta_\beta_n \subset \mathcal{F}_{KL} \times C^\alpha[A, B] \times \mathbb{R}, n \geq 1 \) and constants \( c_1, c_2, c_3, C_1, C_2, C_3 > 0 \) such that

1. \( \sum_{n=1}^{\infty} E_{\Pi_n} \Phi_n < \infty \)

2. \( \sup_{(f_0, v_0) \in U^c \times \Theta_n} E_{\Pi_n} (1 - \Phi_n) \leq C_1 e^{-c_1 n} \)

3. \( \Pi_v(\Theta_v) \leq C_2 e^{-c_2 n}, \Pi_\beta(\Theta_\beta_n) \leq C_3 e^{-c_3 n} \)

4. For all \( \delta > 0 \) and for almost every data sequence \( \{y_i, t_i, x_i\}_{i=1}^{\infty} \),

\[
\Pi\{ (f, \eta) : K_i(f, \eta) < \delta \forall \iota, \sum_{i=1}^{\infty} \frac{V_i(f, \eta)}{\Pi_n(f, \eta)} < \infty \} > 0.
\]

To verify 1 - 4, we construct the sieves \( \Theta_n = U \times \Theta_v \times \Theta_\beta_n \) as \( U \times \{ v : v = \sum_{j=1}^{J} c_j b_j(t), c_j \in [-M_n, M_n], J \leq J_n \} \times \{ \beta : \left| \beta \right| < T_n \} \) for sequences \( J_n, T_n \) to be chosen later. Now we calculate the complement and entropy of \( \Theta_v \) and \( \Theta_\beta_n \). We claim

1. \( \Pi(\Theta_v) \leq \frac{2 J_n (1 + J_n)}{\sqrt{2 \pi M_n}} e^{-M_n^2/2} + P(J \geq J_n), \log N(\epsilon, \Theta_v, \| \cdot \|_\infty) \leq J_n \log (\frac{2M_n}{\epsilon}), \)

2. \( \Pi_\beta(\Theta_\beta_n) \leq \frac{4}{\sqrt{2 \pi M_n}} e^{-T_n^2/2}, \log N(\epsilon, \Theta_\beta_n, \| \cdot \|) \sim o(n) \)

Define \( \Theta_v = \mathcal{F}^v_{J_n} = \{ v : v = \sum_{j=1}^{J} c_j b_j(t), c_j \in [-M_n, M_n], J \leq J_n \} \). First observe that \( P(\text{at least one } c_j \notin [-M_n, M_n]|J) \leq \sum_{j=1}^{J} P(\text{at least one } c_j \notin [-M_n, M_n]|J) \). Without loss of generality, we can assume \( c_j | J \sim N(0, 1) \). Hence \( P(c_j \notin [-M_n, M_n]|J) = \frac{2}{\sqrt{2 \pi}} \int_{-M_n}^{M_n} e^{-c_j^2/2} dc_j = \frac{2}{\sqrt{2 \pi M_n}} e^{-M_n^2/2} \). Thus, \( P(\text{at least one } c_j \notin [-M_n, M_n]|J) \leq \frac{4J_n}{\sqrt{2 \pi M_n}} e^{-M_n^2/2} \). Hence

\[
\Pi(\Theta_v) = \sum_{J=1}^{J_n} P(\text{at least one } c_j \notin [-M_n, M_n]|J) P(J) + P(J \geq J_n),
\]

\[
\leq \sum_{J=1}^{J_n} \frac{4J_n}{\sqrt{2 \pi M_n}} e^{-M_n^2/2} + P(J \geq J_n) = \frac{2J_n (1 + J_n)}{\sqrt{2 \pi M_n}} e^{-M_n^2/2} + P(J \geq J_n)
\]
\( P(J \geq J_n) \) is exponentially small by the hypothesis of the theorem. Choosing \( M_n \sim O(\sqrt{n}) \), the first term is exponentially small. Now, we verify the metric entropy for \( \mathcal{F}^n_{J_n} \). If \( v_1 \) and \( v_2 \) in \( \mathcal{F}^n_{J_n} \), we have \( ||v_1 - v_2||_\infty \leq \sum_{j=1}^{J} |c_j^{(1)} - c_j^{(2)}|B_j(t) \), which implies \( |c_j^{(1)} - c_j^{(2)}| < \epsilon, \forall j \Rightarrow ||v_1 - v_2||_\infty < \epsilon \). Thus, we have \( \epsilon \)-covering number of the set \( \mathcal{F}^n_{J_n} \) with respect to the supremum norm, denoted by \( N(\epsilon, \mathcal{F}^n_{J_n}; || \cdot ||_\infty) < N(\epsilon, [-M_n, M_n]^{J_n}, d) \) and \( \log N(\epsilon, \mathcal{F}^n_{J_n}; || \cdot ||_\infty) < \log(\frac{2M_n}{\epsilon}) \epsilon_n = J_n \log(\frac{2M_n}{\epsilon}) \sim o(n) \) by the hypothesis of the theorem.

Also \( \Theta_{\beta_n} = \{ \beta : |\beta| < T_n \} \), then the \( \log N(\epsilon, \Theta_{\beta_n}, || \cdot ||) \leq \log(\frac{2T_n}{\epsilon}) \). Since \( \beta \) follows Gaussian distribution, \( \Pi_{\beta}(\Theta_{\beta_n}^c) \leq \frac{4}{\sqrt{2\pi T_n}} e^{-T_n^2/2} \). Choosing \( T_n = O(\sqrt{n}) \), \( \Pi_{\beta}(\Theta_{\beta_n}^c) \) can be made exponentially small and \( \log N(\epsilon, \Theta_{\beta_n}, || \cdot ||) \sim o(n) \). Hence condition 3 is satisfied.

**Condition 1 and 2:** We write \( \mathcal{W}_n \) as a disjoint union of three tractable regions. Note that \( U = \mathcal{W}_{1n} \cup \mathcal{W}_{2n} \cup \mathcal{W}_{3n} \), where \( \delta_1 > 0, \delta_2 > 0 \),

\[
\mathcal{W}_{1n} = \mathcal{U}^c \times \{ v : ||v - v_0|| < \delta \} \times \{ |\beta - \beta_0| < \delta_2 \}, \\
\mathcal{W}_{2n} = \{ ||v - v_0||_\infty > \delta_1 \}, \\
\mathcal{W}_{3n} = \{ ||v - v_0||_\infty < \delta_1 \} \times \{ |\beta - \beta_0| > \delta_2 \}.
\]

The proof of the existence of exponentially consistent sequence of tests for \( H_0 : (f, v, \beta) = (f_0, v_0, \beta_0) \) vs \( H_1 : (f, v, \beta) \in \mathcal{W}_{1n} \cap \Theta_n, \mathcal{W}_{2n} \cap \Theta_n \) and \( \mathcal{W}_{3n} \cap \Theta_n \) follows the argument in Proposition 1 of Pati and Dunson (2012) and Proposition 3.1 of Amewou-Atisso et al. (2003) (applied to \( |\beta - \beta_0| > \delta_2 \).

**Condition 4:** This follows trivially from Lemma 1 and noting that \( V(f_0, f_M) \leq \int ||\log f_M - \log f_0||^2 \) is finite as \( ||\log f_M||, ||\log f_M||^2 \) are both bounded by \( f_0 \)-integrable functions.

\[ \square \]

**Remark 4.** If \( J \sim \text{Poiss}(\lambda_J) \), \( P(J \geq J_n \mid \lambda_J) \leq \exp\{-CJ_n \log J_n \} \). Choosing \( J_n = \Omega(n/\log n) \), we can make \( P(J \geq J_n \mid \lambda_J) \) exponentially small.

**Remark 5.** If \( J = ||\tilde{J}|| \), which is the integer closest to \( \tilde{J} \), where \( \tilde{J} \) has Gaussian (or truncated Gaussian) prior. Then \( P(J) = \frac{2}{\sqrt{2\pi}} \int_{-0.5}^{0.5} e^{-J^2/2} dJ \). Hence \( P(J \geq J_n) \) is approximately \( \frac{4}{\sqrt{2\pi J_n}} e^{-J_n^2/2} \). We can choose \( J_n = \Omega(\sqrt{n}) \) to make \( P(J \geq J_n) \) exponentially small.

**Remark 6.** When the smooth \( v(t) \) are combinations of B-splines, we do not need a prior for \( J \) to achieve posterior consistency. We can just let \( \Theta_{\beta n} \) to be \( \mathcal{F}^{J_n}_{J_n} \). Then \( \Pi(\Theta_{\beta n}) \) is exponentially small if \( J_n \sim \Omega(\sqrt{n}) \) and the entropy will be \( o(n) \) if \( M_n \sim \Omega(\sqrt{n}) \).

### 1.4 Simulation Study

In this section, we illustrate the finite sample properties of our methods and compare with the existing methods in a variety of simulation settings. We tabulate results (estimates) obtained from the following classes of error distribution models:
(1) M-estimator using the least absolute deviation loss function of He et al. (2002) (called “M-estimator” model in Table 1.1), (2) the skewed double exponential model of (1.9) for $e_i$ (called “SDE” in Table 1.1), (3) the skewed Gaussian kernel model in (1.21) (called “SPM1” in Table 1.1) and (4) the skewed uniform kernel model in (1.5) (called “SPM2” in Table 1.1). For the skewed Gaussian kernel model, we use

$$f_e(e) = \begin{cases} 
\int_0^\infty \frac{1}{\sqrt{2\pi} \theta} \exp\{-e^2/(2\theta^2)\} dG(\theta) & \text{if } e \geq 0, \\
\int_0^\infty \frac{1}{\sqrt{2\pi} \lambda} \exp\{-e^2/(2\lambda^2 \theta^2)\} dG(\theta) & \text{if } e < 0,
\end{cases}$$

(1.21)

where $G(\theta)$ has the same DP prior as we have mentioned earlier. This skewed Gaussian mixture model is a subclass of skewed uniform kernel mixture model. For the M-estimator in He et al. (2002), we minimise the absolute deviation $\rho(r) = |r|$ using quantile regression with median to be estimated.

To compare the performance of the methods, we used the mean square error $\text{MSE}(\beta) = \sum_{k=1}^s (\hat{\beta}_k - \beta_0)^2 / s$, and $\text{MISE}= \sum_{k=1}^s \sum_{i=1}^n \sum_{j=1}^m (v_0(t_{ij}) - B(t_{ij})^T \hat{\xi}_k)^2 / s$, where $\hat{\beta}_k$ is Bayes estimate of the coefficient vector for $k$th replicated dataset $k, k = 1, \ldots, s$, $\beta_0$ is the true regression coefficient vector and $\hat{\xi}_k$ is the estimator for the spline coefficients for $k$th dataset. The MISE is mean integrated squared error for the nonparametric function estimator which a measure of differences between the estimated and the true function at the observed time points.

Convergence was monitored using the Raftery & Lewis diagnostic test as well as trace plots of the deviance parameters. Also, we get essentially identical posterior summaries with different MCMC starting points and moderate changes to hyperparameters. Each of the simulation settings uses $s = 50$ replicates of simulated data sets, each with 50 subjects measured at 5 time points. The simulation results are listed in Table 1.1.

The first simulation setting is based on a model with common marginal variance. We used $e_{ij} = z_{ij} I(z_{ij} \geq 0) + 2z_{ij} I(z_{ij} < 0)$ and $y_{ij} = v(t_{ij}) + x_{ijp}^T \beta_0 + e_{ij}$, where $z_i \sim \text{MVN}(0, \Sigma)$ with $\Sigma_{ii} = 1$ and $\Sigma_{jk} = \rho = 0.2$ for $j \neq k$. Also, $\beta_0 = (3, -1)'$ and $v(t) = t - t^2$. $x_{ijp}$ are independently sampled from $N(0,1)$ for $p = 1, 2$; $t_{ij}$ are sampled from ordered $U(0,1)$. The prior for each $\xi$ and $\beta_0$ is chosen as $N(0, 1)$, prior for $\rho$ is $U(0, 1)$, prior for $\lambda$ and $\gamma$ are $\exp(1)$. For the skewed Gaussian kernel and skewed uniform kernel models, we took $\alpha = 1$ and $G_0$ as the cdf of $Ga(1, 0.25)$.

Our second simulation setting is based on a model with heteroscedasticity. It is similar to the first simulation setting, except we replace $z_i \sim \text{MVN}(0, \Sigma_i)$, with $\Sigma_i = (x_{i11})^2 \Sigma$. The covariates $x_{ij1} = x_{i11}, j = 1, \ldots, m$ are sampled from $N(0,1)$.

Our third simulation setting aims to evaluate the robustness of the methods to the outliers. The error vectors are sampled from the mixture of two multivariate normal distributions with 95% from $\text{MVN}(0, \Sigma)$ and 5% from the outlier density $\text{MVN}(0, 50\Sigma)$.

In the fourth simulation setting, we simulate observations from the Gaussian copula model with marginal density as the Gumbel distribution with location $-1$ and
Table 1.1: Approximate (via Monte Carlo) sampling mean and mean square error (MSE within parenthesis) of the estimated regression parameters ($\beta_1, \beta_2$); and approximate mean integrated squared error (MISE) for function $v(t)$. True parameter values are $\beta_1 = 3, \beta_2 = -1$.

<table>
<thead>
<tr>
<th>Simulation study 1: Homoscedastic</th>
<th>$\beta_1$(100×MSE)</th>
<th>$\beta_2$(100×MSE)</th>
<th>MISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-estimator</td>
<td>2.99(1.496)</td>
<td>-1.02(1.193)</td>
<td>18.780</td>
</tr>
<tr>
<td>SDE</td>
<td>2.98(0.499)</td>
<td>-1.01(0.642)</td>
<td>12.310</td>
</tr>
<tr>
<td>SPM1</td>
<td>2.99(0.438)</td>
<td>-1.00(0.456)</td>
<td>7.274</td>
</tr>
<tr>
<td>SPM2</td>
<td>3.00(0.629)</td>
<td>-1.00(0.514)</td>
<td>9.717</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation study 2: Heteroscedastic</th>
<th>$\beta_1$(100×MSE)</th>
<th>$\beta_2$(100×MSE)</th>
<th>MISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-estimator</td>
<td>3.05(2.901)</td>
<td>-1.00(0.271)</td>
<td>3.086</td>
</tr>
<tr>
<td>SDE</td>
<td>3.00(1.775)</td>
<td>-1.00(0.111)</td>
<td>3.095</td>
</tr>
<tr>
<td>SPM1</td>
<td>2.99(2.938)</td>
<td>-1.00(0.089)</td>
<td>2.100</td>
</tr>
<tr>
<td>SPM2</td>
<td>2.99(4.678)</td>
<td>-1.00(0.099)</td>
<td>2.374</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation study 3: Outlier</th>
<th>$\beta_1$(100×MSE)</th>
<th>$\beta_2$(100×MSE)</th>
<th>MISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-estimator</td>
<td>3.00(0.736)</td>
<td>-1.00(0.820)</td>
<td>8.270</td>
</tr>
<tr>
<td>SDE</td>
<td>2.99(0.648)</td>
<td>-1.00(0.654)</td>
<td>5.945</td>
</tr>
<tr>
<td>SPM1</td>
<td>2.99(0.380)</td>
<td>-1.00(0.495)</td>
<td>5.152</td>
</tr>
<tr>
<td>SPM2</td>
<td>2.98(0.472)</td>
<td>-1.00(0.587)</td>
<td>6.577</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation study 4: Gumbel</th>
<th>$\beta_1$(100×MSE)</th>
<th>$\beta_2$(100×MSE)</th>
<th>MISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-estimator</td>
<td>2.98(7.264)</td>
<td>-1.04(6.360)</td>
<td>71.49</td>
</tr>
<tr>
<td>SDE</td>
<td>2.91(5.158)</td>
<td>-0.96(4.620)</td>
<td>69.860</td>
</tr>
<tr>
<td>SPM1</td>
<td>2.93(4.143)</td>
<td>-0.94(3.068)</td>
<td>44.410</td>
</tr>
<tr>
<td>SPM2</td>
<td>2.94(4.744)</td>
<td>-0.94(3.461)</td>
<td>42.800</td>
</tr>
</tbody>
</table>

scale $-1/\log(\log(2))$. Both the third and fourth simulation studies aim to study the robustness of our methods to the misspecification of error distribution in (1.5).

Our $v(t)$ is a quadratic function, and we use quadratic splines for all the models and the simulation settings. For the $M$-estimator, we select the knots using the BIC methods described in He et al. (2002). In simulations 1 and 4, we use fixed knots at $(0.25, 0.375, 0.50, 0.625, 0.75)$ and in simulations 2 and 3, we use knots at $(0.33, 0.67)$.

It can be seen from Table 1.1 that all our median regression models outperform the $M$-estimators for the common marginal variance set up in simulation study 1 with respect to the MSE of parameters and MISE of the unknown smooth function. In simulation study 2, somewhat surprisingly, the estimates based on our median regression models are even robust to the heteroscedastic error of the simulation model. The skewed double exponential error model outperforms the $M$-estimator in estimating
the parameter (comparable smooth function estimations) in the heteroscedastic set up 3, while the skewed Gaussian kernel and skewed uniform kernel models perform better in estimating $\beta_2$ and the smooth function compared to the $M$-estimator. In both simulation settings 3 and 4, all our models outperform $M$-estimator with respect to the MSE of parameters and the MISE of unknown smooth function.

### 1.5 Analysis of P2C2 Study

We illustrate the use of our estimators for the analysis of a longitudinal study designed to monitor heart disease and the progression of cardiac abnormalities in children born to HIV-infected women Lipshultz et al. (1998). The Pediatric Pulmonary and Cardiac Complications (in short, P2C2) has been previously analyzed by Lipsitz et al. (2009) and Parzen et al. (2011) after dichotomizing the time between heartbeats (RRBO). As discussed in the introduction, the RRBO is related to cardiac health and is measured via an EKG. However, no one has presented an analysis of the continuous version of RRBO, which is known to be highly skewed. Here, our goal is to estimate the median regression function of the RRBO. For the sake of brevity, we omit the exploratory plots and figures showing the non-linear time effect, high skewness and stationarity of the error distribution.

In P2C2 study, a birth cohort of 432 infants born to women infected with HIV have RRBO measured approximately every half year from birth to up to 7 years old. The longitudinal response of interest is the RRBO. There are 2007 observations in total for all the 432 children. Different children are tested at different time points. Also the observations of each child are unbalanced. There are 56 children with only one observation, and 2 children with 16 observations (which is the maximum number of observations).

The HIV status of the child (1=Yes, 0=No) is the main covariate of interest. Of the 432 children involved in this study, 356(82.41%) are HIV negative, 76(17.59%) are HIV positive. The HIV status does not change over time for the same child. The main purpose of our analysis is to develop a suitable model of the effects of HIV and age (time variable) on highly skewed longitudinal response RRBO, estimate the regression parameter corresponding to the HIV status effect and make prediction of future RRBO of the same as well as new (similar) children.

We use the quantiles (0.25, 0.375, 0.50, 0.625, 0.75) of age as the fixed knots to approximate the $v(t)$ by $B(t) = [b_1(t), ..., b_8(t)]$ using cubic B-splines method. The model is

$$y_{ij} = \beta_1 + \beta_2 HIV_i + B(t_{ij})^T \xi + e_{ij}.$$ 

Based on variogram plots in Figure 1.1 and 1.2, to model the Gaussian copula of (1.7), we use an uniform correlation structure for HIV negative group and an exponential correlation structure for HIV positive group.

The longitudinal continuous response RRBO is right-skewed, and we expect the distribution of skewed errors are different for HIV positive and negative children.
The errors are expected to have median 0. We used the M-estimator as in He et al. (2002), a skewed double exponential model, and a skewed Gaussian kernel model. For all models, the median regression function is the same.

For the parametric skewed double exponential model: \( f_e(e) \) is defined as

\[
    f_e(e) = \begin{cases} 
        \gamma \exp\{-\gamma e\}/2 & \text{if } e \geq 0, \\
        \gamma \exp\{\gamma e/\lambda_0\}/(2\lambda_0) & \text{if } e < 0 \text{ & } HIV = 0, \\
        \gamma \exp\{\gamma e/\lambda_1\}/(2\lambda_1) & \text{if } e < 0 \text{ & } HIV = 1,
    \end{cases}
\]

where \( \lambda_0, \lambda_1 \) are positive parameters.
For the semiparametric skewed Gaussian kernel model: $f_e(e)$ is defined as

$$f_e(e) = \begin{cases} 
\int_0^\infty \frac{1}{\sqrt{2\pi}\theta} \exp\left\{-\frac{e^2}{2\theta^2}\right\} dG(\theta) & \text{if } e \geq 0, \\
\int_0^\infty \frac{1}{\sqrt{2\pi}\lambda_0\theta} \exp\left\{-\frac{e^2}{2\lambda_0^2\theta^2}\right\} dG(\theta) & \text{if } e < 0 \text{ and } HIV = 0, \\
\int_0^\infty \frac{1}{\sqrt{2\pi}\lambda_1\theta} \exp\left\{-\frac{e^2}{2\lambda_1^2\theta^2}\right\} dG(\theta) & \text{if } e < 0 \text{ and } HIV = 1, 
\end{cases}$$

A Bayesian MCMC method is used in these two models, with the prior for all $\xi$ equal to $N(0, 1)$; the prior for $\beta_1$ and $\beta_2$ equal to $N(0, 1)$, the prior for $\lambda_0, \lambda_1$ equal to $\exp(1)$; the prior for $\rho_0, \rho_1$ equal to $U(0, 1)$ in the covariance matrix $\Sigma$. The form of $\Sigma$ is similar to the following matrix $\Sigma_0 = \begin{pmatrix} 1 & \rho_0 \\
\rho_0 & 1 \end{pmatrix}$, and $\Sigma_1 = \begin{pmatrix} 1 & \rho_1^{t_{i2} - t_{i1}} \\
\rho_1^{t_{i2} - t_{i1}} & 1 \end{pmatrix}$. 

Figure 1.2: Variogram Plot: HIV=1
One of the major advantages of our Bayes method is the ease of prior specifications (based on prior opinions about median). We use exp(1) prior for $\gamma$ for the model with skewed double exponential error. For specifying the prior for the nonparametric $F_e$ in (1.5), we assume $G_0$ follows a gamma distribution with shape 1 and scale 0.33 for a skewed Gaussian kernel model. For the $M$-estimator, we perform quantile regression to minimise the least absolute deviation loss function and use BIC for knots selection. Results can be seen in Table 1.2. The median of RRBO from our model seems less negatively affected by HIV status than the $M$-estimator. However, our median regression models have smaller standard deviations than the $M$-estimator.

Table 1.2: Results of the analysis of HIV study with the estimators and the corresponding standard errors (within parenthesis) of the regression parameters ($\beta_1, \beta_2$) and other model parameters under different methods.

<table>
<thead>
<tr>
<th>Estimation models</th>
<th>SDE</th>
<th>SPM1</th>
<th>$M$-estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>Estimate (Sd)</td>
<td>Estimate (Sd)</td>
<td>Estimate (Sd)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.440 (0.005)</td>
<td>0.448 (0.006)</td>
<td>0.428 (0.038)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.041 (0.0039)</td>
<td>-0.041 (0.0046)</td>
<td>-0.042 (0.0063)</td>
</tr>
<tr>
<td>$\lambda_0$</td>
<td>0.715 (0.043)</td>
<td>0.749 (0.037)</td>
<td>.</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>0.839 (0.068)</td>
<td>0.864 (0.058)</td>
<td>.</td>
</tr>
<tr>
<td>$\rho_0$</td>
<td>0.234 (0.028)</td>
<td>0.242 (0.028)</td>
<td>.</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td>0.168 (0.031)</td>
<td>0.192 (0.035)</td>
<td>.</td>
</tr>
</tbody>
</table>

1.6 Conclusions

In this paper, we have developed a novel semi-parametric Bayes model and associated method for longitudinal data with skewed continuous responses. Our method has the advantage of handling stationarity of marginal distribution, a nonparametric effect of time (and some other covariate), ease of physical interpretation of the regression parameter and feasibility of implementation via MCMC tools. Our MCMC based analysis is implementable even via freely available software such as WINBUGS.

In our finite-sample simulation study, our method outperforms existing methods assuming either symmetric or skewed parametric residual distributions. We show (via simulation studies presented in Table 1.1) that our Bayes likelihood method gives a much lower MSE for a finite sample compared that of $M$-estimations in He et al. (2002) for common marginal variance data. For the first time, for likelihood and Bayes methods for skewed longitudinal data, we also provide theoretical support for the proposed methodology which guarantees consistent estimates of the regression coefficients under mild conditions. Our results on consistency are useful to even derive the rate of posterior convergence of the regression coefficients. However, for the sake
of brevity we omit the proof of the optimal rate of convergence of our semiparametric Bayes estimators.

It would be interesting to theoretically quantify the finite-sample gain in using a skewed median zero residual density by comparing the mean squared error of our method to that using a symmetric residual density. Another possible methodological direction is to develop computationally efficient heteroscedastic median zero residual distributions to accommodate outliers with large errors and obtain robust regression estimates. However, these are topics of possible future research are beyond the scope of this present work.
CHAPTER 2

BAYESIAN VARIABLE SELECTION FOR SKEWED HETEROSCEDASTIC ERROR

2.1 Introduction

In this article, we consider a general linear regression model with \( n \) observations on a dependent variable \( Y \) and \( p \) predictors, \( y = X\beta + \epsilon \), where \( y \) = \((y_1, \ldots, y_n)^T\) is the response vector, \( X \in \mathbb{R}^{n \times p} \) is the \( n \times p \) fixed matrix of predictors, \( \beta \in \mathbb{R}^p \) is a fixed (unknown) coefficient vector and \( \epsilon \) is the \( n \times 1 \) vector of skewed and heteroscedastic errors. Our motivating example is from the Medical Expenditure Panel Survey (MEPS) (Cohen, 2003; Natarajan et al., 2008) for the year 2002, which was conducted by the United States National Center for Health Statistics, Centers for Disease Control and Prevention. Medical cost data are typically highly skewed to the right, in that a small percentage of subjects sustain extremely high costs compared with other subjects. With such data, the variance of the outcome tends to increase as the mean increases.

For the iid normal errors \( \epsilon \) with mean 0 and variance \( \sigma^2 \), the Lasso of Tibshirani (1996) estimates linear regression coefficients through \( L_1 \)-constrained least squares. Lasso estimates often are viewed as \( L_1 \)-penalized least squares estimates. They minimize the residual sum of square with a constraint which is expressed in term of \( L_1 \)-norm of \( \beta \) is \( \beta_{\text{Lasso}} = \arg\min_{\beta} (y - X\beta)^T(y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \) for some \( \lambda \geq 0 \). When \( \lambda = 0 \), \( \beta_{\text{Lasso}} \) is the ordinary least-square estimands, and it will shrink to 0 when \( \lambda \) is efficiently large. The entire path of Lasso estimates for all values of \( \lambda \) can be efficiently computed through a modification of the LARS algorithm of Efron et al. (2004). Tibshirani (1996) suggests that the Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical laplace priors. Motivated by that connection, Park and Casella (2008) consider a fully Bayesian analysis using a conditional Laplace prior specification of the form \( \pi(\beta | \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda |\beta_j| / \sqrt{\sigma^2}} \) and the noninformative scale-invariant marginal
prior $\pi(\sigma^2) = 1/\sigma^2$ on $\sigma^2$. Later, Kyung et al. (2010) apply hierarchical models and Gibbs sampling to get estimators and valid Bayesian standard errors for generalized lasso estimators. However, the double exponential prior for parameters can not get exact zeros. Meanwhile, our data example does not satisfy the iid normal assumption.

When the normality or symmetric assumption of $\epsilon$ does not hold, the lasso estimators are not appropriate. For example, our medical cost data are highly skewed. Penalized median 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 $\beta_{\text{Lasso}} = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - X_i \beta) + \lambda \sum_{j=1}^{p} |\beta_j|$ (Koenker, 2005), with $\tau = 0.5$ for median regression, $\rho_\tau$ is the loss function

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

However, with our medical cost data, the variance of the outcome tends to increase as the mean increase. The outcome variable is the subject’s ‘total health care expenditures in the year 2002’. Previous analyses (Natarajan et al., 2008) of medical cost data suggest that the variance is approximately equal to the mean raised to the power of 1.5, which is non-standard variance function. Unlike the lasso or penalized quantile regression methods, we propose a likelihood based method which can handle the skewness and heteroscedasticity. Our estimators produce exact zeros for the regression parameters. We first provide a new class of prior which can select the sparse variables and estimate the parameters simultaneously. Our models accommodate the situation when the median as well as the scale and shape of response distribution are affected by the covariates. The parameters of median regression can be estimated via Hierarchical model after applying the generalized Box-Cox transformation in Bickel and Doksum (1981) to both the response variable and the $X\beta$ predictor term. We assume the transformed response have a Gaussian or symmetric distribution with transformed $X\beta$ as the mean and median. However, outliers may exist even after the generalized Box-Cox transformation to contaminate the Gaussian assumption. We propose two novel models to handle the outliers or the heavy tail distribution. The prior for the regression parameter will change according to the transformation function.

In this paper, our aim is to develop a robust Bayesian variable selection and parameter estimation procedure in median regression using the Transform-both-sides model. The rest of the article proceeds as follows. In section 2, we propose our new prior. The transform-both-sides model will also be proposed in section 2. Section 3 shows a medical cost data example using our method. Section 4 presents the simulation study results. Section 5 concludes with some discussions and possible directions for future research.
2.2 Transform-both-sides Model and Prior

2.2.1 Transform-both-sides model

In this section, we propose the transform-both-sides model to handle the skewed heteroscedastic error. We assume that after an appropriate monotone transformation, the error will be Gaussian or Symmetric. Bickel and Doksum (1981) propose a monotone power transformation, an extension of the Box-Cox power family,

\[ g_\eta(y) = \frac{\text{sgn}(y)|y|^{\eta - 1}}{\eta}, \eta > 0 \]

where

\[ \text{sgn}(y) = \begin{cases} 1, & \text{if } y \geq 0 \\ -1, & \text{otherwise} \end{cases} \]

The transformation \( g_\eta(y) \) is monotone with positive derivative \( g'_\eta(y) = |y|^{\eta - 1} \). We assume that under an optimal \( \eta \), the transformed response \( g_\eta(y) \) has a symmetric and unimodal distribution. We also restrict the \( \eta \in (0, 2) \). Follow this, the transform-both-sides model becomes

\[ g_\eta(y_i) = g_\eta(X_i \beta) + e_i \quad (2.1) \]

where \( e_i \) are iid errors with cumulative density function \( F_e \) and variance \( \sigma^2 \). Thus, \( g_\eta(y) \) is supposed to be symmetric and unimodal with mean and median \( g_\eta(X_i \beta) \) with optimal \( \eta \). The approximate variance of \( y_i \) is \( \sigma^2|X_i \beta|^{2-2\eta} \). This allows our model to be useful for dealing with heteroscedasticity of \( y \). In this paper, we assume \( F_e \) is Gaussian \( N(0, \sigma^2) \) distribution, or Gaussian distribution with some outliers or some heavier tail distribution. Both of our two models follow this assumption. The likelihood for \( g_\eta(y_i) \) is \( f_e(g_\eta(y_i) - g_\eta(X_i \beta)) \), where \( f_e \) is the unimodal symmetric density function. Thus, the likelihood for \( y_i \) is \( f_e(g_\eta(y_i) - g_\eta(X_i \beta))g'_\eta(y_i) \). For the none-outlier case, we assume \( e_i \) follows the Gaussian distribution with \( \phi(e) \) as the probability density function.

From frequentist point of view, a ridge penalty is desired to account for noise and collinearity in the data, but it never encourages sparsity. Thus, to maintain accuracy and promote sparsity, one must take into account nonconvex penalties such as those of type \( l_0 + l_2^2 \). She (2009, 2011) illustrate the good properties of this type of penalties in detail. Here, from Bayesian point of view, we specify a spike Gaussian prior to mimic this nonconvex penalty. In the meanwhile, the estimations from Bayesian MCMC will achieve the global optimal point as it has integration instead of differentiation. We propose a Spike Gaussian prior for \( \beta \) with probability \( \tilde{\pi} \) at point 0, with probability \( 1 - \tilde{\pi} \) from Gaussian distribution with mean 0 and variance \( \sigma_\beta^2 \). This prior can give exact zeros for the sparse regression parameters. With probability \( 1 - \tilde{\pi} \), the parameters have the Gaussian prior and will give estimation for non-zero coefficients. The following, we derive the relation between this Spike Gaussian prior with the \( l_0 + l_2 \)
penalty. Our prior has the form

\[ \pi(\beta) = \begin{cases} \tilde{\pi}\delta_0, & \text{if } \beta = 0, \\ (1 - \tilde{\pi})\phi(\beta, 0, \sigma_\beta^2), & \text{otherwise}, \end{cases} \]

where \( \phi(\beta, 0, \sigma_\beta^2) \) is the pdf for the normal density with mean 0 and standard deviation \( \sigma_\beta \). Thus, we have

\[ P(\beta_j \leq x_j, x_j < 0) = (1 - \tilde{\pi})\Phi(x_j), \]
\[ P(\beta_j \leq x_j, x_j \geq 0) = (1 - \tilde{\pi})\Phi(x_j) + \tilde{\pi}1_{(x_j \geq 0)}, \]
\[ f_{\beta_j}(x_j) = (1 - \tilde{\pi})\phi(x_j) + \tilde{\pi}\delta_0(x_j), \]
\[ = (1 - \tilde{\pi})\exp\{-x_j^2/(2\sigma^2)\}/\sqrt{2\pi\sigma^2} + \tilde{\pi}\exp\{-x_j^2/a^2\}/\sqrt{\pi a^2}, \]

where \( a \to 0 \),

\[ \prod_{j=1}^p f_{\beta_j}(x_j) = \prod_{j=1}^p [\frac{(1 - \tilde{\pi})\phi(x_j) + \tilde{\pi}\delta_0(x_j)}{\sqrt{2\pi\sigma^2}}], \]
\[ \log \prod_{j=1}^p f_{\beta_j}(x_j) = \sum_{j=1}^p \log\left[ (1 - \tilde{\pi})\phi(x_j) + \tilde{\pi}\delta_0(x_j) \right], \]
\[ = \sum_{j \in \text{nz}} \log\left[ (1 - \tilde{\pi})\phi(x_j) \right] + \sum_{j \in \text{z}} \log\left[ (1 - \tilde{\pi})\phi(0) + \tilde{\pi}\delta_0(0) \right], \]
\[ = \sum_{j \in \text{nz}} \left[ \log(1 - \tilde{\pi}) - x_j^2/(2\sigma^2) - 0.5\log(2\pi\sigma^2) \right] \]
\[ + \sum_{j \in \text{z}} \log\left[ (1 - \tilde{\pi})\sqrt{2\pi\sigma^2} + \tilde{\pi}\delta_0(0) \right], \]
\[ = ||x||_0 \log\left[ (1 - \tilde{\pi})/\sqrt{2\pi\sigma^2} \right] - ||x||_2^2/(2\sigma^2) \]
\[ + (p - ||x||_0)\log\left[ (1 - \tilde{\pi})/\sqrt{2\pi\sigma^2} + \tilde{\pi}\delta_0(0) \right], \]
\[ = -||x||_2^2/(2\sigma^2) - ||x||_0 \log\frac{(1 - \tilde{\pi})/\sqrt{2\pi\sigma^2} + \tilde{\pi}\delta_0(0)}{(1 - \tilde{\pi})/\sqrt{2\pi\sigma^2}} \]
\[ + p\log\left[ (1 - \tilde{\pi})/\sqrt{2\pi\sigma^2} + \tilde{\pi}\delta_0(0) \right]. \]

Our Spike Gaussian prior is related to the \( l_0 + l_2 \) penalty function, which can result in good variable selection and parameter estimation. Meanwhile, we can get a prior guess for \( \sigma_\beta^2 \), variance in the sparse parameter prior. For large \( n \) case, the ordinary least square method can be applied to get \( \beta_{\text{ols}} \). From the estimated \( \beta_{\text{ols}} \), we can derive the prior guess for \( \sigma_\beta^2 \).

However, the Gaussian or Symmetric error is derived after transform-both-sides model. The parameter \( \eta \) in the monotone transformation function takes an very
important role on the shape and scale of the distribution. We know that the variance of $y$ can be approximated by $\sigma^2 |X_i \beta|^{2-2\eta}$. When $\eta > 1$, the variance will be small when the regression function is greater than 1. When $\eta < 1$, the variance of $y$ will be large when the regression function is large, the shape of $y$ will be very hard to catch. Thus, we specify the prior for $\beta$ dependent on $\eta$. When one $x_{ij} \beta_j$ is larger than the others, $g_\eta(X_i \beta) \approx g_\eta(x_{ij}) g_\eta(\beta_j)$. When several terms are comparable, $g_\eta(X_i \beta)$ can be approximated by the scaled $g_\eta(x_{ij}) g_\eta(\beta_j)$. From this point of view, instead of specifying the prior for $\beta$ and $\eta$ independently, we specify the prior for $g_\eta(\beta)$ as spike Gaussian. This will result in another spike prior for $\beta$ given $\eta$.

$$P(\beta \leq \beta_0) = P(g_\eta(\beta) \leq g_\eta(\beta_0)),$$

$$= \pi_1 g_\eta(\beta_0) + (1 - \tilde{\pi}) \Phi \sigma_\beta^2 (g_\eta(\beta_0)),$$

$$f_\beta(\beta|\eta) = \pi_1 \delta_0 + (1 - \tilde{\pi}) \phi(g_\eta(\beta_0), 0, \sigma_\beta^2) g'_\eta(\beta_0).$$

For the none-outlier data, the Bayesian hierarchical model is represented as

$$g_\eta(y_i) \sim N(g_\eta(x_i \beta), \sigma^2), i = 1, \ldots, n,$$

$$y_i|\beta, \gamma, \sigma^2, \eta \sim \phi(g_\eta(y_i) - g_\eta(x_i \beta), \sigma^2) g_\eta(y_i),$$

$$\beta|\eta \sim \prod_{j=1}^p f_\beta(\beta_j|\eta),$$

$$\sigma^2 \sim IG(a, b),$$

$$\eta/2 \sim Beta(c, d),$$

### 2.2.2 Outlier or heavy tail distribution

After generalized Box-Cox monotone transformation, the normal assumption may be violated by outliers. Hampel et al. (1986) estimate that a routine data set may contain about $1\% - 10\%$ (or more) outliers. Follow this, our first model becomes

$$g_\eta(y_i) = g_\eta(X_i \beta) + \gamma_i + e_i$$

where $e_i$ are iid $N(0, \sigma^2)$ error. $\gamma_i$ is nonzero if $i$ observation is an outlier, $\gamma_i$ is zero if $i$ observation is not an outlier. This formulation is applied by She and Owen (2011), and earlier applied by McCann and Welsch (2007). This mean shift model allows any combination of observations to be outliers. It has $n + p$ regression parameters and only $n$ data points.

We assume that after transformation, $g_\eta(X_i \beta)$ is the mean and median of $g_\eta(y_i)$. This leads that the median of $\gamma_i + e_i$ is zero. From this point of view, we assume $\gamma_i$ has a symmetric distribution (see Appendix). As $\gamma$ is sparse, we apply the similar Spike
Gaussian prior on $\gamma$. The prior is $f_\gamma(\gamma) = \pi_\gamma \delta_0 + (1 - \pi_\gamma) \phi(\gamma, 0, \sigma_\gamma^2)$. The resultant Bayesian process will be

$$g_\eta(y_i) \sim N(g_\eta(x_i \beta) + \gamma_i, \sigma^2), i = 1, \ldots, n,$$

$$y_i | \beta, \gamma, \sigma^2, \eta \sim \phi(g_\eta(y_i) - g_\eta(X_i \beta) - \gamma_i, \sigma^2) g_\eta'(y_i),$$

$$\beta | \eta \sim \prod_{j=1}^p f_\beta(\beta_j | \eta),$$

$$\gamma \sim \prod_{i=1}^n f_\gamma(\gamma_i),$$

$$\sigma^2 \sim IG(a, b),$$

$$\eta/2 \sim Beta(c, d),$$

Lange and Sinsheimer (1993) propose the normal independent (NI) distributions family. An element of the NI family is defined as the distribution of the $p$-variate random vector $y = \mu + U^{-1/2}Z$, where $\mu$ is a location vector, $Z$ is a normal random vector with mean vector 0, variance-covariance matrix $\Sigma$ and $U$ is a mixing positive random variable with cumulative distribution function $H(u | \nu)$ and probability density function (pdf) $h(u | \nu)$, independent of $Z$. Lachos et al. (2011) state that the NI family constitutes a class of thick-tailed distributions, some of which are the multivariate versions of the Student’s-t, slash and contaminated normal (CN) distributions if the error is multivariate normal. For our model, we apply this NI family to the univariate case. The hierarchical process is

$$g_\eta(y_i) = g_\eta(X_i \beta) + u_i^{-1/2} e_i$$

where $e_i$ are iid $N(0, \sigma^2)$

$$g_\eta(y_i) \sim N(g_\eta(x_i \beta), U_i^{-1} \sigma^2), i = 1, \ldots, n$$

$$y_i | \beta, \gamma, \sigma^2, \eta \sim \phi(g_\eta(y_i) - g_\eta(X_i \beta), U_i^{-1} \sigma^2) g_\eta'(y_i)$$

$$U_i \sim H(\cdot | \nu)$$

$$\beta | \eta \sim \prod_{j=1}^p f_j(\beta_j)$$

$$\sigma^2 \sim IG(a, b)$$

$$\eta/2 \sim Beta(c, d)$$

where $f(\beta)$ is the spike double gamma density as defined in the previous subsection. And we will present the results of three kinds of heavier tail distribution (Student’s-t, slash and contaminated normal) in our simulation study in section 3. Now, we briefly
introduce these three distributions (Lange and Sinsheimer, 1993) and what priors will be used for \( \nu \) (Lachos et al., 2011).

The \( t \) distribution is recovered by taking the scale variable \( U \) to have a \( \chi^2/\nu \) distribution with \( \nu > 0 \) possibly noninteger. In this case the distribution function \( H(u) \) has density

\[
    h(u) = \frac{(\nu/2)^{\nu/2}u^{\nu/2-1}e^{-u/2}}{\Gamma(\nu/2)}
\]

We set the degrees of freedom parameter \( \nu \) has a truncated exponential prior distribution on the interval \((2, \infty)\).

The slash distribution has scale variable \( U \) with density \( h(u) = \nu u^{\nu-1} \) on \([0, 1]\). And a \( \text{Gamma}(a, b) \) distribution with small positive values of \( a \) and \( b(b \ll a) \) is adopted as a prior distribution for \( \nu \).

For the contaminated normal distribution, the scale variable \( U \) is concentrated at the two points \( \lambda < 1 \) and 1 with masses \( \phi \) and \( 1 - \phi \). A \( \text{Beta}(\nu_0, \nu_1) \) distribution is used as a prior for \( \nu \), and an independent \( \text{Beta}(\rho_0, \rho_1) \) is adopted as prior for \( \rho \).

### 2.3 Data Example

Our data example comes from the Medical Expenditure Panel Survey (Cohen, 2003; Natarajan et al., 2008), the outcome variable is the subject’s ‘total health care expenditures in the year 2002’. Previous analyses of medical cost data suggest that the variance is approximately equal to the mean raised to the power of 1.5, which is a non-standard variance function. Medical cost data are typically highly skewed to the right, in that a small percentage of subjects sustain extremely high costs compared to other subjects (see Figure 2.1). With such data, the variance of the outcome tends to increase as the mean increases. Meanwhile we have 24 independent variables including age, gender, race, disease history, etc. However, from the previous analyses of the cost data, only a few variables contribute to the response variable cost.

We picked up one geographical region and one ‘primary sampling units’ (PSUs) within stratum as our data example in order to avoid the sample weight for different strata and different units. In this data, we have 215 patients in total (with 6.5% patients having no cost at all, and maximum $130869), however, some independent variables have missing values. Thus, after dropping out the missing values, we have 173 patients left. There are 24 independent variables. The minimum cost is 0 and the maximum is $79660, with a mean $4584 and median $1342. We analyse this data set by our proposed model and the penalized quantile regression model.

For computational reason, we standardize the response cost variable, the AGE02X, HIDEGYR, RTHLTH53, BMINDX53 and ADOVER42 while keeping all the other variables as what they are (binary variables with values 0 and 1). From our proposed model, we detect ADOVER42 (-0.1639739), stroke (0.9023082) and need_med (-0.3357590) are significant effects to the cost data. While penalized quantile regres-
Figure 2.1: Histogram of Medical Cost for Training Data

In order to better understand the pattern of our dataset, we first want to verify whether our data set satisfy the Gaussian assumption. We calculate the $U^T y$ (where $X = UDV$ is the SVD for the design matrix $X$) to see if it satisfy the normal assumption. From Figure 2.2, we can see that the data does not satisfy the normal assumption.

Then, we would like to know whether the penalized quantile regression will perform...
better than our transform-both-sides model. For each method, we calculate Loss2 and Loss3. Loss2 is derived from the prediction error $y - X^T\hat{\beta}$. We assume this prediction error follows a double exponential distribution, transform this prediction error to the standard normal distribution, and plot the QQ-plot and histogram plot. Loss3 is derived by $g_\hat{\eta}(y) - g_\hat{\eta}(X^T\hat{\beta}) - \hat{\gamma}$. In order to avoid the inaccuracy of estimated $\gamma$, we just keep these subjects which are not outliers, and then plot the QQ-plot and histogram plot. See Fig 2.3 and 2.4. We can find out that after transformation of
both sides, the loss is closer to the normal distribution than the other two losses.

Figure 2.3: QQ plot and Histogram for Loss2 and Loss3: Penalized Quantile

The test data is from one ‘primary sampling units’ of another geographical region. In this data set, there are 26 patients in total. The minimum medical cost is 0 and maximum cost is $7492, with mean $1313 and median $728. The medical cost is also right skewed (see Figure 2.5). The sum of prediction error square $\sum_{i=1}^{26}(y_i - \hat{y_i})^2$ from our model is 1.956429 (1.896624 for $\sum_{i=1}^{26}(y_i - \hat{y_i})$, where $\hat{y_i}$ is the median of MCMC posterior samples). The penalized quantile regression gives us 2.281046 for
the sum of prediction error square. Our model presents less prediction error than the penalized quantile regression method.

In order to better understand the prediction of test data. We first plot the empirical CDF for the test data (see the circle points in Figure 2.6), then for each posterior sample of predicted value (bootstrap method is applied for the penalized quantile regression to get the $X\hat{\beta}$ as the sampled predicted value) for the response in test data, we calculate the empirical CDF. For each patient in test data, we get
the middle 95%, 99.5% and 100% values of empirical CDF, and we treat the lower and upper value of empirical CDF as the corresponding confidence interval for the observed empirical CDF. The graphs for both the penalized quantile regression and our transform-both-sides model are listed in Figure 2.6, 2.7 and 2.8.

We can see that our transform-both-sides value perform very well to cover the true empirical CDF in the 95%, 99.5% and 100% confidence interval except when the response value are relatively small. Our model also gives a broader interval than the
Figure 2.6: Empirical CDF 95%
Figure 2.7: Empirical CDF 99.5%
Empirical CDF vs observed y: 100%

Figure 2.8: Empirical CDF 100%
penalized quantile regression method.

2.4 Simulation Studies

In this section, we introduce our simulation studies to evaluate the performance of our proposed models. The Lasso Penalized Quantile Regression in Koenker (2005) is applied for comparison. We simulate each model 50 times. We list the simulation results using the following measures in She and Owen (2011):

- **M**: the mean masking probability (fraction of undetected true outliers),
- **S**: the mean swamping probability (fraction of good points labeled as outliers),
- **JD**: the joint outlier detection rate (fraction of simulations with 0 masking).

For the penalized quantile regression, the program does not provide exact zeros, we set up a cutoff point 0.01 to distinct a value from zero. The tuning parameter $\lambda$ was selected via a grid search based on the cross-validation tuning error in terms of the check loss function $\rho_\tau(y - X\beta)$ evaluated on the test data. 5 folds are applied for the cross-validation.

2.4.1 None-outlier

For none outlier simulations, we simulated different scenarios to test whether our methods perform better than quantile lasso. For each scenario, there are 50 simulations with sample size 50. We calculate the average number of nonzeros, masking rate, swamping rate and joint detection rate to compare the performance. We also compare the estimation of parameters by $L/L^* - 1$, where

$$L = \frac{1}{n} \sum_{i=1}^{n} (g_{\eta'}(y_i) - g_{\eta'}(X_i\hat{\beta}))^2/(2\sigma^2) - n/2\log(2\pi\sigma^2) + (\eta^* - 1) \sum_{i=1}^{n} \log(|y_i|),$$

$$L^* = \frac{1}{n} \sum_{i=1}^{n} (g_{\eta'}(y_i) - g_{\eta'}(X_i\beta^*))^2/(2\sigma^2) - n/2\log(2\pi\sigma^2) + (\eta^* - 1) \sum_{i=1}^{n} \log(|y_i|),$$

$L$ is the log-likelihood under the estimated parameters $\hat{\beta}$ and $L^*$ is the log-likelihood under the true parameters. For all the scenarios, the data are sampled from the model

$$g_\eta(y) = g_\eta(X\beta) + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, and we set $\sigma^2 = 1$. We also fix $a = 2, b = 2, c = 1, d = 1$ for the prior. This way, the prior for $\eta$ is $U(0, 2)$.

Scenario 1: $\eta^* = 0.5, p = 8, \beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0).$
Scenario 2: $\eta^* = 1.8, p = 8, \beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0).$

Form the scenario 1 and 2, we can see that for the $L/L^* - 1$, both our models and penalized quantile regression model give a decent estimation, however, the penalized
### Table 2.1: Results for simulation study: Scenario 1.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Ave $L/L^* - 1$</th>
<th>Ave no. of nz</th>
<th>M(%)</th>
<th>S(%)</th>
<th>JD(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penalized Quantile</td>
<td>0.036230</td>
<td>5.84</td>
<td>0</td>
<td>56.8</td>
<td>100</td>
</tr>
<tr>
<td>Spike Gaussian</td>
<td>-0.015070</td>
<td>3.16</td>
<td>0</td>
<td>3.2</td>
<td>100</td>
</tr>
<tr>
<td>NI-t</td>
<td>-0.014710</td>
<td>3.16</td>
<td>0</td>
<td>3.2</td>
<td>100</td>
</tr>
<tr>
<td>NI-slash</td>
<td>-0.016060</td>
<td>3.16</td>
<td>0</td>
<td>3.2</td>
<td>100</td>
</tr>
<tr>
<td>NI-CN</td>
<td>-0.015370</td>
<td>3.14</td>
<td>0</td>
<td>2.8</td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 2.2: Results for simulation study: Scenario 2.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Ave $L/L^* - 1$</th>
<th>Ave no. of nz</th>
<th>M(%)</th>
<th>S(%)</th>
<th>JD(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penalized Quantile</td>
<td>0.037040</td>
<td>5.82</td>
<td>0</td>
<td>56.4</td>
<td>100</td>
</tr>
<tr>
<td>Spike Gaussian</td>
<td>-0.057430</td>
<td>3.02</td>
<td>0</td>
<td>0.4</td>
<td>100</td>
</tr>
<tr>
<td>NI-t</td>
<td>-0.052600</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>NI-slash</td>
<td>-0.056160</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>NI-CN</td>
<td>-0.05594</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

Quantile model give much variant $L/L^* - 1$ with six number summary
$(0.073840, 0.013420, 0.001364, 0.036230, 0.045610, 0.427800)$ when $\eta^* = 0.5$
and $(-0.199500, -0.057010, -0.008905, 0.037040, 0.113800, 0.694700)$ when $\eta^* = 1.8$.
In the mean time, our Spike Gaussian model has
$(-0.105000, -0.023640, -0.009702, -0.015070, -0.001523, 0.031110)$ for $\eta^* = 0.5$
and $(-0.228100, -0.076870, -0.034620, -0.057430, -0.022410, 0.003255)$ for $\eta^* = 1.8$.

Comparing the average number of non-zeros and swamping, all our models perform
much better than the penalized quantile regression. Both our models and penalized
quantile regression perform very good at masking and joint detection.

We found out that when $\eta^* = 0.5$, our models perform worse than $\eta^* = 1.8$
comparing the average number of non-zeros and swamping. Thus, the following sim-
ulations, we just compare our first model with the penalized quantile regression model
under different cases for $\eta^* = 0.5$ and $\eta^* = 1.8$. Instead of using $p = 8$, we increase
our $p = 20$ and we set number of non-zeros equals 12.

Case 1: $\beta^* = (2, \ldots, 2, 0, \ldots, 0)$

Case 2: $\beta^* = (-10, \ldots, -10, 4, \ldots, 4, 0, \ldots, 0)$

Case 3: $\beta^* = (-10, \ldots, -10, 4, 4, 0, \ldots, 0)$

Case 4: $\beta^* = (-10, -10, -4, -4, -2, -2, 2, 4, 4, 10, 10, 0, \ldots, 0)$
Case 5: $\beta^* = (-10, \ldots, -10, 2, \ldots, 2)$

Case 6: $\beta^* = (-10, -10, -8, -8, -6, -6, -4, -4, -2, -2, 2, 0, \ldots, 0)$

From these simulation study results in Table 2.3, we can clearly see the effect of $\eta^*$ on the performance of our models. When the ratio of the maximum signal and minimum signal is large and when there are many small clusters, both our model and penalized quantile model do not perform very good. For all the cases, our model perform better when $\eta^* = 1.8$ than $\eta^* = 0.5$ comparing average number of non-zeros, masking, swamping and joint detection. The same pattern happens for penalized quantile regression method.

2.4.2 Heavy tail distribution

For outlier simulations, there is one extra table for the $\gamma$ only for spike Gaussian model. Our NI model does not provide the estimation of outliers, it provide the robust estimation of $\beta$ estimates. We simulated different scenarios to test whether our methods perform better than quantile lasso. For each scenario, there are 50 simulations with sample size 50. We calculate the average number of non-zeros, masking rate, swamping rate and joint detection rate to compare the performance. We also compare the estimation of parameters by $L/L^* - 1$, where

$$L = \sum_{i=1}^{n} (g_{\eta^*}(y_i) - g_{\eta^*}(X_i\hat{\beta}) - \gamma_i^*)^2/(2\sigma^2) - n/2\log(2\pi\sigma^2) + (\eta^* - 1)\sum_{i=1}^{n} \log(|y_i|),$$

$$L^* = \sum_{i=1}^{n} (g_{\eta^*}(y_i) - g_{\eta^*}(X_i\beta^*) - \gamma_i^*)^2/(2\sigma^2) - n/2\log(2\pi\sigma^2) + (\eta^* - 1)\sum_{i=1}^{n} \log(|y_i|),$$

$L$ is the log-likelihood under the estimated parameters $\hat{\beta}$ and $L^*$ is the log-likelihood under the true parameters. For all the scenarios, the data are sampled from the model

$$g_\eta(y) = g_\eta(X\beta) + \gamma + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, and we set $\sigma^2 = 1$. We also fix $a = 2, b = 2, c = 1, d = 1$ for the prior. This way, the prior for $\eta$ is $U(0, 2)$.

Scenario 3: $\eta^* = 0.5, p = 8, \gamma_{(1:2)} = 8, \gamma_3 = -8, \gamma_{(4:50)} = 0, \beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$. From the result, for the spike Gaussian model, we have average 3.32 non-zeros, the masking is 1.33%, swamping is 0.765974% and joint detection is 98%.

Scenario 4: $\eta^* = 1.8, p = 8, \gamma_{(1:2)} = 8, \gamma_3 = -8, \gamma_{(4:50)} = 0, \beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$. From the result, for the spike Gaussian model, we have average 3.28 non-zeros for $\gamma$, the masking is 0%, swamping is 0.5957% and joint detection is 100%.

The NI family produce similar results as the spike Gaussian family. Both of our two models have better results than the penalized quantile regression in $L/L^* - 1$,
Table 2.3: Comparison of Spike Gaussian model and Penalized Quantile model.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Spike Gaussian</th>
<th>Penalized Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>η* = 0.5</td>
<td>η* = 1.8</td>
<td>η* = 0.5</td>
</tr>
<tr>
<td>Case 1 Ave. (L/L^* - 1)</td>
<td>-0.046840</td>
<td>2.7190</td>
</tr>
<tr>
<td>Ave. no. non-zeros</td>
<td>13.6</td>
<td>12.02</td>
</tr>
<tr>
<td>M (%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S(%)</td>
<td>20</td>
<td>0.25</td>
</tr>
<tr>
<td>JD(%)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Case 2 Ave. (L/L^* - 1)</td>
<td>-0.02612</td>
<td>0.0336600</td>
</tr>
<tr>
<td>Ave. no. non-zeros</td>
<td>13.26</td>
<td>12</td>
</tr>
<tr>
<td>M (%)</td>
<td>0.67</td>
<td>0</td>
</tr>
<tr>
<td>S(%)</td>
<td>16.75</td>
<td>0</td>
</tr>
<tr>
<td>JD(%)</td>
<td>94</td>
<td>100</td>
</tr>
<tr>
<td>Case 3 Ave. (L/L^* - 1)</td>
<td>-0.02644</td>
<td>-0.058300</td>
</tr>
<tr>
<td>Ave. no. non-zeros</td>
<td>12.94</td>
<td>12</td>
</tr>
<tr>
<td>M (%)</td>
<td>0.33</td>
<td>0</td>
</tr>
<tr>
<td>S(%)</td>
<td>12.25</td>
<td>0</td>
</tr>
<tr>
<td>JD(%)</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>Case 4 Ave. (L/L^* - 1)</td>
<td>-0.011100</td>
<td>0.08554</td>
</tr>
<tr>
<td>Ave. no. non-zeros</td>
<td>12.58</td>
<td>12</td>
</tr>
<tr>
<td>M (%)</td>
<td>6.17</td>
<td>0</td>
</tr>
<tr>
<td>S(%)</td>
<td>16.5</td>
<td>0</td>
</tr>
<tr>
<td>JD(%)</td>
<td>46</td>
<td>100</td>
</tr>
<tr>
<td>Case 5 Ave. (L/L^* - 1)</td>
<td>0.007530</td>
<td>0.040650</td>
</tr>
<tr>
<td>Ave. no. non-zeros</td>
<td>11.32</td>
<td>12</td>
</tr>
<tr>
<td>M (%)</td>
<td>11.17</td>
<td>0</td>
</tr>
<tr>
<td>S(%)</td>
<td>8.25</td>
<td>0</td>
</tr>
<tr>
<td>JD(%)</td>
<td>24</td>
<td>100</td>
</tr>
<tr>
<td>Case 6 Ave. (L/L^* - 1)</td>
<td>-0.011820</td>
<td>0.09576</td>
</tr>
<tr>
<td>Ave. no. non-zeros</td>
<td>12.1</td>
<td>12</td>
</tr>
<tr>
<td>M (%)</td>
<td>6.67</td>
<td>0</td>
</tr>
<tr>
<td>S(%)</td>
<td>11.25</td>
<td>0</td>
</tr>
<tr>
<td>JD(%)</td>
<td>44</td>
<td>100</td>
</tr>
</tbody>
</table>

average number of non-zeros, masking, swamping and joint detection. Our models can deliver a robust estimator when the normal assumption failed. From our simulation studies, η plays an important role for the variable selection and coefficient estimation.
Table 2.4: Results for simulation study: Scenario 3.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Ave $L/L^* - 1$</th>
<th>Ave no. of nz</th>
<th>M(%)</th>
<th>S(%)</th>
<th>JD(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penalized Quantile</td>
<td>1.0370</td>
<td>5.68</td>
<td>0</td>
<td>53.6</td>
<td>100</td>
</tr>
<tr>
<td>Spike Gaussian</td>
<td>-0.007251</td>
<td>3.22</td>
<td>0.67</td>
<td>4.8</td>
<td>98</td>
</tr>
<tr>
<td>NI-t</td>
<td>-0.009156</td>
<td>3.18</td>
<td>0</td>
<td>3.6</td>
<td>100</td>
</tr>
<tr>
<td>NI-slash</td>
<td>-0.006310</td>
<td>3.16</td>
<td>0</td>
<td>3.2</td>
<td>100</td>
</tr>
<tr>
<td>NI-CN</td>
<td>-0.009884</td>
<td>3.2</td>
<td>0</td>
<td>4</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2.5: Results for simulation study: Scenario 4.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Ave $L/L^* - 1$</th>
<th>Ave no. of nz</th>
<th>M(%)</th>
<th>S(%)</th>
<th>JD(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penalized Quantile</td>
<td>3.544</td>
<td>5.54</td>
<td>0</td>
<td>50.8</td>
<td>100</td>
</tr>
<tr>
<td>Spike Gaussian</td>
<td>-0.04564</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>NI-t</td>
<td>-0.04290</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>NI-slash</td>
<td>-0.039280</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>NI-CN</td>
<td>-0.04461</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

2.5 Discussion

In this chapter, we developed a new method via transform-both-sides model to simultaneously perform variable selection and parameter estimation. Our proposed two novel models can deal with skewed and heteroscedastic response. Our outlier detection model can estimate the regression parameters and detect outliers successfully and jointly. We found that our model work better than penalized quantile regression method under different simulation scenarios. The parameter $\eta$ plays an important role in controlling the shape of the data set. Meanwhile, it is critical for the model performance. Our model can handle the $p > n$ case, only the computation will take longer time. Further investigation on high-dimensional problems and the non-asymptotic will be conducted in the future.
CHAPTER 3

PRACTICAL SEMIPARAMETRIC BAYES ANALYSIS OF HETEROSCEDASTIC AND SKEWED RESPONSE

3.1 Introduction

In many epidemiological studies, often the main goal is to assess the changes in a response distribution for various levels of exposure, while adjusting for the effects of additional covariates. For example, Longnecker et al. (2001) studied the association between the exposure variable, DDE concentration in mother’s serum during the third trimester of pregnancy, and the response variable, gestational age at delivery (in short, GAD). The study also included various additional covariates including the demographic variables such as age. There is a vast literature on frequentist as well as Bayesian semiparametric regression methods for the continuous response (e.g., Robinson, 1988; Ruppert et al., 2003). These models often impose the assumption such as common variance and/or symmetric unimodal error distribution. However, such modeling restrictions are often inappropriate for real-life studies including this GAD study because the exposure variable may affect the location as well as the shape and scale of the skewed response distribution. For heteroscedastic response, a standard practice, particularly in reproductive epidemiology (e.g., Longnecker et al., 2001, and the references therein) is to dichotomize the response variable followed by a binary regression analysis. However, the choice of the cut-off point may not be obvious, and the resulting method may cause a substantial loss of information about the relationship between the response and the exposure. From public health perspective, it is of interest to assess the effects of DDE exposure on the entire left tail (lower quantiles) of the GAD distribution, because the effect of DDE on the earlier gestation period is believed to be more important than its effect on the later gestation period (Dunson and Park, 2008).
There is a long history of Bayesian methods for density estimation using a kernel mixture model with an unknown mixing distribution that is assigned a prior $\Pi$. The most common choice of prior $\Pi$ is the Dirichlet process (Ferguson, 1973, 1974). Recently, many have focused on generalizing the mixture model to the density regression setting in which the conditional density of $Y$ given $x$ changes flexibly is modelled nonparametrically (for example, to Müller et al., 1996; Griffin and Steel, 2006, 2008; Dunson et al., 2007; Dunson and Park, 2008; Chung and Dunson, 2009; Tokdar et al., 2010; Pati et al., 2013, among others). However, it is difficult to interpret the effect of DDE on GAD from a non-parametric density of $Y$ given $x$.

We aim to combine the physical interpretation of semiparametric regression with the modeling flexibility of the Bayesian density estimation for heteroscedastic response. To achieve this, some (e.g., Kottas and Gelfand, 2001; Lin et al., 2012; Hanson and Johnson, 2002) have considered nonparametric residual distribution in regression. These approaches often have the unappealing property of discontinuous residual density (e.g., Kottas and Gelfand, 2001). Some of these methods do not allow heteroscedasticity (e.g., Hanson and Johnson, 2002; Kottas and Gelfand, 2001) and others assume the variance to be a parametric function of the median (e.g., Lin et al., 2012). In addition, these approaches lead to highly discretized estimates of the residual density particularly for moderate sample sizes. Lavine and Mockus (1995) only allow a single predictor and an unknown residual density with a monotonicity constraint. Pelenis (2012) in an unpublished technical report considers residual densities with some constraints on moments.

There are several practical and relevant issues that provide motivation for this article. First, to simplify inferences and prior determination, it is important to have a class of residual densities allowing interpretation through quantiles of the observable responses. Existing methods based on discrete mixtures have difficulty in centering the prior on an initial guess (a possibly parametric density). Moreover, priors for residual densities based on discrete mixtures (e.g., Pati and Dunson, 2013; Pelenis, 2012) often require the estimation of infinitely many stochastic processes. This is highly computationally demanding proposition irrespective of the dimension of the covariate space.

Inclusion of program as a supplement, our focus is on developing a semiparametric regression model which not only possesses good theoretical properties and physical interpretation, but is also easy to implement using existing packages (such as Matlab). Our method is computationally efficient, simple in structure, and allows easy incorporation of prior information. In addition, our method allow multiple modes for the density function, which is very popular in the real-life data sets.

This paper is organized as follows. We provide our model and prior specifications in Section 3.2. In this section, we also characterize the support of the prior demonstrating the flexibility and scope of the class of the priors. Section 3 presents the likelihood, posterior and associated details for the Bayesian computation via MCMC. The methods are illustrated via several simulation examples in Section 3.4. We discuss
the application to the GAD study in Section 3.5. Proofs of the results are deferred to the Appendix.

### 3.2 Model and Prior Specification

Let \( Y_i \) and \( x_i \) be respectively the continuous response variable and the \( p \)-dimensional covariate corresponding to the subject (patient) \( i = 1, \ldots, n \). Without loss of generality assume that \( x_i \in \mathcal{X} = [0, 1]^p \). Let \( X = [x_1 : x_2 : \cdots : x_n]^T \) be the \( n \times p \) covariate matrix. The regression model is expressed as

\[
Y_i = h(x_i) + \epsilon_i, \quad \text{for} \quad \epsilon_i \sim f(\cdot | x_i),
\]

where the heteroscedastic residual density \( f(\cdot | x_i) \) has location (mean) 0, but, with shape and scale dependent on covariate \( x_i \).

For ease of interpretation of the covariate effects, we will consider a linear \( h(x_i) = x_i \beta \), although extension to either a semiparametric (e.g., Yau et al.; Chan et al., 2006) or to a completely nonparametric case (Pati and Dunson, 2013) is straightforward. To ensure the interpretation and identifiability of the vector of regression parameters \( \beta \), we assume \( \int_{-\infty}^{+\infty} \epsilon f(\epsilon | x)de = 0 \). Next, we state our prior distribution for the collection of residual densities induced by predictors \( \{f(\cdot | x) : x \in \mathcal{X}\} \). As mentioned before, the assumption of symmetric error is inappropriate for epidemiological study (as well as in other studies). Also, it is often difficult to determine a suitable transformation of the response to achieve a symmetric error. For example, in our reproductive study, the gestational age at delivery (GAD) is skewed at the left tail and the skewness may depend on the predictors. From previous studies (e.g., Longnecker et al., 2001; Dunson and Park, 2008), we expect to have multiple modes with the primary mode near the mean. Therefore it is important to develop a class of prior for the residual density \( f(\cdot | x) \) to allow a wide family of skewed and heteroscedastic residual densities.

#### 3.2.1 Model for heteroscedastic residual density

We create a novel class of residual density functions which are restricted to have mean zero, but avoids the discrete mixture formulation in illustrated in Pelenis (2012). One specified case for our model is Kundu and Dunson (2011), where the non-linear latent factor model was motivated by machine learning methods for non-linear dimensionality reduction (Bishop and Tipping, 1998; Lawrence, 2005).

Let the latent variable residual density (LV-RD) model be specified by

\[
\epsilon_i \mid x_i, \eta_i \sim \tilde{\pi} N\left\{ \mu(x_i, \eta_i), \sigma^2 \right\} + (1 - \tilde{\pi}) N\left\{ \frac{-\tilde{\pi} \mu(x_i, \eta_i)}{1 - \tilde{\pi}}, \sigma^2 \right\},
\]

\[
\mu \sim \Pi_\mu, \quad \tilde{\pi} \sim \Pi_{\tilde{\pi}}, \quad \sigma \sim \Pi_\sigma, \quad \eta_i \sim U(0, 1),
\]

where \( i = 1, \ldots, n \), \( \eta_i \in [0, 1] \) are subject-specific latent variables, \( \mu \in C([0, 1]^p \times [0, 1]) \) is a transfer function relating the latent variables and the covariates to the residuals.
\[\epsilon_i \text{ and } \tilde{\pi} \in [0, 1]. \tilde{\pi} \text{ is a parameter with support } [0, 1]. \text{ It takes an important role in adjusting the skewness and heteroscedasticity. If } \tilde{\pi} = 0.5, \text{ we will get a symmetric probability density function. When } \tilde{\pi} \to 0 \text{ or } \tilde{\pi} \to 1, \text{ the density will be more asymmetric. } \Pi_\mu \text{ is a prior on } C[0, 1]^{p+1}, \Pi_\tilde{\pi} \text{ is a prior on } [0, 1] \text{ and } \Pi_\sigma \text{ is a prior on } [0, \infty).

The density of } \epsilon \text{ given } x \text{ conditional on transfer function } \mu \text{ and scale } \sigma \text{ is obtained on marginalizing out the latent variable } \text{as } f_x(\epsilon) := f_{\mu,\sigma,\tilde{\pi}}(\epsilon \mid x) =
\[\int_0^1 \tilde{\pi} \phi[\epsilon - \mu(x, \eta)] + (1 - \tilde{\pi}) \phi[\epsilon - \frac{\tilde{\pi} \mu(x, \eta)}{1 - \tilde{\pi}}] \, d\eta. \tag{3.4}\]

Clearly, the conditional mean } E(\epsilon \mid x, \eta) = \tilde{\pi} \mu(x, \eta) + (1 - \tilde{\pi}) \frac{\tilde{\pi} \mu(x, \eta)}{1 - \tilde{\pi}} = 0 \text{ implying } \int_{-\infty}^{\infty} \epsilon f_x(\epsilon) \, d\epsilon = 0. \text{ Let } \mathcal{F}_r \text{ denote the space of all residual densities with mean 0.}
\[
\mathcal{F}_r = \left\{ f(\epsilon \mid x) = h(x, \epsilon), h : [0, 1]^p \times \mathbb{R} \to \mathbb{R}^+, \int_{\mathbb{R}} h(\epsilon, x) \, d\epsilon = 1, \int_{\mathbb{R}} \epsilon h(x, \epsilon) \, d\epsilon = 0 \forall x \in \mathbb{R}^p \right\}
\]

and } C([0, 1]^{p+1}) \text{ denotes the space of all continuous functions on } [0, 1]^{p+1}. \text{ Define a map } g : C([0, 1]^{p+1}) \times [0, \infty) \to \mathcal{F}_r \text{ with } g(\mu, \sigma) = f_{\mu,\sigma,\tilde{\pi}}. \text{ One can induce a prior } \Pi \text{ on } \mathcal{F}_r \text{ via the mapping } g \text{ by placing independent priors } \Pi_\mu \text{ and } \Pi_\sigma \text{ on } C([0, 1]^{p+1}) \text{ and } [0, \infty) \text{ respectively, with } \Pi = (\Pi_\mu \otimes \Pi_\sigma \otimes \Pi_{\tilde{\pi}}) \circ g^{-1}. \text{ We describe choices for } \Pi_\mu \text{ in the following section.}

### 3.2.2 Choice of } \Pi_\mu

Clearly, the specification (3.4) is not complete unless we place a computationally convenient as well as a flexible prior distribution for \mu. \text{ Typical choices for } \Pi_\mu \text{ include a Gaussian process (GP) prior on } [0, 1]^{p+1}, \text{ tensor-product of splines or radial basis functions. In the later case, one can choose } \mu(x, \eta) = \sum_{j,k} b_{jk} B_j^p(x_i) B_k^1(\eta_i) \text{ where } \{B_j^s, j = 1, \ldots, J\} \text{ for any positive integer } s \text{ denotes the radial basis function on } [0, 1]^s, \text{ whence } B_j^s(t) = e^{-\kappa \|t - t_j\|^2} \text{ for some bandwidth } \kappa > 0 \text{ and knot-points } t_j \in [0, 1]^s.

Figure 3.1 illustrates prior realizations of residual density functions } f(\epsilon \mid x) \text{ for different values of } x. \text{ Here } \mu(x, \eta) = \sum_{j,k} b_{jk} B_j^p(x_i) B_k^1(\eta_i). \text{ It is evident that the resulting residual density functions are asymmetric and multimodal, with degree of skewness and number of modes varying with } x.

### 3.2.3 Prior flexibility

It is not immediately clear whether the class of mean zero residual densities } f_{\mu,\sigma,\tilde{\pi}} \text{ in (3.4) in the range of } g \text{ theoretically encompass a large subset of } \mathcal{F}_r. \text{ We provide an intuition that relates the above class with convolutions and is used in Proposition 1.
Figure 3.1: Density Plot of Skewed and Heteroscedastic Error

to characterize the support of $\Pi_\mu$. Define $\mathcal{F}_{r0}$ such that any $f_0 \in \mathcal{F}_{r0}$ is a continuous mean zero residual density such that

$$
f_0(\epsilon \mid x) = \tilde{\pi}_0 f_0^*(\epsilon \mid x) + \frac{(1 - \tilde{\pi}_0)^2}{\tilde{\pi}_0} f_0^* \left( \frac{-(1 - \tilde{\pi}_0)}{\tilde{\pi}_0} \epsilon \mid x \right),
$$

for an arbitrary conditional density $f_0^*$ not necessarily mean 0 and for some $\tilde{\pi}_0 \in [0, 1]$. This class of probability density functions is large enough to cover the skewed and heteroscedastic density functions. It can generate a density function with one big mode and some small modes at the tail. This is very common in practice. Let the cumulative distribution function of $f_0^*$ be $F_0^*(\epsilon \mid x) = \int_{-\infty}^{\epsilon} f_0^*(z \mid x) dz$. Assume $f_0^*$ to be non-zero almost everywhere within its support, so that $F_0^*(\cdot \mid x) : \text{supp}(f_0^*(\cdot \mid x))$
\( f_{\mu_0, \sigma, \tilde{\pi}_0}(\epsilon \mid \mathbf{x}) = \tilde{\pi}_0 \int_{-\infty}^{\infty} \phi_\sigma(\epsilon - z)f_0^* (z \mid \mathbf{x})dz 
+ \frac{(1 - \tilde{\pi}_0)^2}{\tilde{\pi}} \int_{-\infty}^{\infty} \phi_\sigma(\epsilon - z)f_0^* \left( \frac{- (1 - \tilde{\pi}_0)}{\tilde{\pi}_0} z \mid \mathbf{x} \right)dz \) \hspace{1cm} (3.5)
\( \rightarrow \tilde{\pi}_0 f_0^*(\epsilon \mid \mathbf{x}) + \frac{(1 - \tilde{\pi}_0)^2}{\tilde{\pi}} f_0^* \left( \frac{- (1 - \tilde{\pi}_0)}{\tilde{\pi}_0} \epsilon \mid \mathbf{x} \right) \) \hspace{1cm} (3.6)

as \( \sigma \to 0 \), where the second equality follows from the change of variable theorem.

We have used the following fact to proceed from the second last sentence to the last sentence. \( f_{\mu_0, \sigma, \tilde{\pi}_0}(y \mid \mathbf{x}) = \phi_\sigma \ast f_0^*(y \mid \mathbf{x}) \), the convolution of \( f_0^* \) with a normal density having mean 0 and standard deviation \( \sigma \). It is well known that the convolution \( \phi_\sigma \ast f_0^*(\cdot \mid \mathbf{x}) \) can approximate \( f_0^*(\cdot \mid \mathbf{x}) \) for each \( \mathbf{x} \) arbitrary closely as the bandwidth \( \sigma \to 0 \). More precisely, for \( f_0^*(\cdot \mid \mathbf{x}) \in L^p(\lambda) \) for any \( p \geq 1 \) for each \( \mathbf{x} \), \( \| \phi_\sigma \ast f_0^*(\cdot \mid \mathbf{x}) - f_0^*(\cdot \mid \mathbf{x}) \|_{p,\lambda} \to 0 \) as \( \sigma \to 0 \). Furthermore, a stronger result \( \| \phi_\sigma \ast f_0^*(\cdot \mid \mathbf{x}) - f_0^*(\cdot \mid \mathbf{x}) \|_\infty = O(\sigma^2) \) for each \( \mathbf{x} \) holds if \( f_0^*(\cdot \mid \mathbf{x}) \) for each \( \mathbf{x} \) is compactly supported. Assume \( \mathbf{x}_i \)'s are drawn from a probability measure \( Q \) on \([0, 1]^p\) with a density \( q \), define the \( L_1 \) support of a prior \( \Pi \) on \( \mathcal{F}_r \) to be the subset

\[
\left\{ f \in \mathcal{F}_r : \Pi (g \in \mathcal{F}_r : \| f - g \|_{1,q} < \epsilon \text{ for any } \epsilon > 0) \right\}
\]

where \( \| f - g \|_{1,q} = \int_{[0,1]^p} \int_y |g(y \mid \mathbf{x}) - f(y \mid \mathbf{x})|dyq(\mathbf{x})d\mathbf{x} \).

The following Proposition 1 characterizes the \( L_1 \) support of the LV-RD prior \( \Pi \) in terms of a subset of \( \mathcal{F}_{r0} \).

**Proposition 1.**

1. \( \mathcal{F}_{r0} \subset \mathcal{F}_r \).

2. Assume \( \Pi_\mu \) has full sup-norm support on \( C[0, 1]^{p+1} \), i.e., given any \( \epsilon > 0 \), \( \Pi_\mu(\| \mu - \mu_0 \|_\infty < \epsilon) > 0 \) for \( \mu_0 \in C[0, 1]^{p+1} \). Also assume \( \Pi_\sigma \) and \( \Pi_{\tilde{\pi}} \) has full support on \([0, \infty)\) and \([0, 1]\) respectively. Then the \( L_1 \) support of the induced prior \( \Pi \) on \( \mathcal{F}_{r0} \) contains

\[
\left\{ f_0 \in \mathcal{F}_{r0} : \int_{[0,1]^p} \int_y |y|f_0^*(y \mid \mathbf{x})q(\mathbf{x})dyd\mathbf{x} < \infty \right\}.
\]

i.e. all residual densities which have a finite first moment and are non-zero almost everywhere on their support.
3.2.4 Analogy with discrete mixture models

Let \( \tilde{\lambda} \) denote the Lebesgue measure on \([0,1]\), or equivalently, the \( U[0,1] \) distribution and \( B(\mathbb{R}) \) the set of all Borel subsets of \( \mathbb{R} \). For any measurable function \( \mu : [0,1]^{p+1} \rightarrow \mathbb{R} \), define for a fixed \( B \in B(\mathbb{R}) \) and \( x \in [0,1]^p \), a function \( \nu_\mu : B(\mathbb{R}) \times [0,1]^p \rightarrow [0,1] \) given by \( \nu_\mu(B;x) := \tilde{\lambda} \{ \mu^{-1}(B;x) \} \), where \( \mu^{-1}(B;x) = \{ y \in [0,1] : \mu(y,x) \in B \} \). Clearly, \( \nu_\mu(B;x) \) is a probability measure for each fixed \( x \), and \( \nu_\mu(B,:) \) is \( \{ \mathbb{R}^p, B(\mathbb{R}) \} \rightarrow \{ \mathbb{R}, B(\mathbb{R}) \} \) measurable function for any Borel set \( B \in B(\mathbb{R}) \). Thus \( \nu_\mu(\cdot;: \cdot) \) defines a valid conditional probability kernel on \( \{ \mathbb{R}, B(\mathbb{R}) \} \).

By the change of variable theorem for induced measures for a fixed \( x \in [0,1]^p \), we can write \( f_{\mu,\sigma,\tilde{\pi}}(\epsilon \mid x) \)

\[
= \int_0^1 f(\epsilon \mid x, \eta) d\eta \\
= \int_0^1 \left[ \tilde{\pi} \phi_\sigma(\epsilon - \mu(x, \eta)) + (1 - \tilde{\pi}) \phi_\sigma \left( \epsilon + \frac{\tilde{\pi} \mu(x, \eta)}{1 - \tilde{\pi}} \right) \right] d\eta, \\
= \int_{-\infty}^{\infty} \left[ \tilde{\pi} \phi_\sigma(\epsilon - t) + (1 - \tilde{\pi}) \phi_\sigma \left( \epsilon + \frac{\tilde{\pi}}{1 - \tilde{\pi}} t \right) \right] \nu_\mu(dt; x),
\]

so that \( f_{\mu,\sigma,\tilde{\pi}} \) can be expressed as a kernel mixture form with conditional mixing measure \( \nu_\mu(\cdot; x) \). It turns out that this mechanism of creating random distributions is very general. Depending on the choice of \( \mu \), one can create a large variety of predictor-dependent conditional mixture distributions based on this specification. For example, if for fixed \( x, y \mapsto \mu(x, y) \) is a strictly monotone function, then \( \nu_\mu(\cdot; x) \) is absolutely continuous with respect to the Lebesgue measure, while choosing \( \mu(\cdot; x) \) to be a step function, one obtains a discrete mixing distribution. Then the resulting residual density will be exactly same as in Pelenis (2012). However, it is easier to place a prior on \( \mu \) supported on the space of continuous functions \( C[0,1]^{p+1} \) without further shape restrictions.

### 3.3 Posterior Computation

Since posterior computation in LV-RD model is not standard, we will outline the steps briefly. Let \( \mu(x, \eta) = \sum_{j,k} b_{jk} B_j^p(x_i) B_k^1(\eta_i) \) where \( B_j^p \) are chosen to be radial basis functions. Also, \( \tau_1 = 1/\sigma_1^2, \tau_2 = 1/\sigma_2^2 \), the parameters are \( \theta = (\tilde{\pi}, \beta, \tau, \{ b_{jk} \}) \).

The likelihood for this model will be

\[
L = \prod_{i=1}^n \left[ \frac{\tilde{\pi} \tau_1^{1/2}}{\sqrt{2\pi}} \exp \left\{ \frac{-(y_i - x_i^\prime \beta - \mu(x_i, \eta_i))^2 \tau_1}{2} \right\} \\
+ \frac{(1 - \tilde{\pi}) \tau_2^{1/2}}{\sqrt{2\pi}} \exp \left\{ \frac{-(y_i - x_i^\prime \beta + \tilde{\pi}/(1 - \tilde{\pi}) \mu(x_i, \eta_i))^2 \tau_2}{2} \right\} \right]
\]
We introduce indicator random variable $z_i$'s to facilitate the posterior computation of $\beta$.

$$z_i = \begin{cases} 
1 & \text{if } y_i \sim N(\mu(x_i, \eta_i), \sigma_1^2) \\
0 & \text{if } y_i \sim N(-\bar{\pi}\mu(x_i, \eta_i)/(1 - \bar{\pi}), \sigma_2^2) 
\end{cases}$$

Let $y = (y_1, \ldots, y_n)^T$, $x = \{x_1, \ldots, x_n\}$, $z = (z_1, \ldots, z_n)^T$, $b = \{b_{jk}\}$ and $\eta = (\eta_1, \ldots, \eta_n)^T$. The joint density of the observed data $y$, $x$ and $\eta$ and unobserved data $z$ are

$$f(\{y, z \mid x, \eta, \theta\}) \propto f(z \mid \theta) f(y \mid x, \eta, z, \theta)$$

From the definition of $z_i$, we have $f(z \mid \theta) = (\bar{\pi})^{-z} (1 - \bar{\pi})^{1-z}$. Meanwhile, we have $f(y_i \mid x_i, \eta_i, z_i, \theta) =$

$$\left[ N(\mu(x_i, \eta_i), \sigma_1^2) \right]^{z_i} \left[ N(-\bar{\pi}\mu(x_i, \eta_i)/(1 - \bar{\pi}), \sigma_2^2) \right]^{1-z_i}.$$

We have

$$f(y, z, \theta, x, u) \propto \pi(\theta) \prod_{i=1}^{n} [f(y_i \mid x_i, \eta_i, z_i, \theta) f(z_i \mid \theta)],$$

where $\pi(\theta)$ is the prior for $\theta$. Choose a $N(0, 1)$ prior for $\beta$, a $\text{Ga}(k_1, k_2)$ prior for $\tau$, where $k_1$ is the shape parameter and $k_2$ is the scale parameter. Also, we choose $\text{N}(0, \psi_{jk})$ prior for each of $b_{jk}$ where $\psi_{jk} = \psi_j \tau_k$, $\psi_j, \tau_k \sim \text{IG}(\nu/2, \nu/2)$. Also, let $\bar{\pi} \sim \text{Beta}(a_1, a_2)$.

The posterior distributions from these conjugate priors can be calculated. For $i = 1, \ldots, n$, let $\mu_i = \mu(x_i, \eta_i)$. The posterior distribution of $z_i$ will be a bernoulli distribution. We have $z_i \mid y, \theta \sim \text{Ber}(\bar{\pi}N(y_i - x_i^T \beta - \mu_i, 0, \sigma_1^2)/[\bar{\pi}N(y_i - x_i^T \beta - \mu_i, 0, \sigma_1^2) + (1 - \bar{\pi})N(y_i - x_i^T \beta + \bar{\pi}/(1 - \bar{\pi})\mu_i, 0, \sigma_2^2)])$. The posterior distribution of $\beta$ is Normal, $\beta_k \mid y, \tau, b, z \sim \text{N}(\mu_{\beta}, \sigma_{\beta}^2)$, with $1/\sigma_{\beta}^2 = \tau_1 \sum_i x_{ik}^2 z_i + 1 + \tau_2 \sum_i x_{ik}^2 (1 - z_i)$ and $\mu_\beta = \left[ \sum_i \tau_1 z_i x_{ik}(y_i - \sum_{j \neq k} x_{ij} \beta_j - \mu_i) + \sum_i \tau_2 (1 - z_i) x_{ik}(y_i - \sum_{j \neq k} x_{ij} \beta_j + \bar{\pi} \mu_i/(1 - \bar{\pi})) \right] \sigma_{\beta}^2$.

$$\tau_1 \sim \text{Ga}(\sum_i z_i / 2 + k_1, 1 / [\sum_i z_i (y_i - x_i^T \beta - \mu_i)^2 / 2 + 1 / k_2]),$$

$$\tau_2 \sim \text{Ga}((n - \sum_i z_i) / 2 + k_3, 1 / [\sum_i (1 - z_i) (y_i - x_i^T \beta + \bar{\pi}/(1 - \bar{\pi})\mu_i)^2 / 2 + 1 / k_4]).$$

$$b_{jk} \mid z, \bar{\pi}, \tau \sim \text{N}(\mu_b, \sigma_b^2),$$

the mean parameter $\mu_b = \left[ \sum B_j^0(x_i) B_k^1(\eta_i) z_i (y_i - x_i^T \beta - (\mu_i - b_{jk} B_j^0(x_i) B_k^1(\eta_i))) / (1 - \bar{\pi}) \sum B_j^0(x_i) B_k^1(\eta_i) (1 - z_i) (y_i - x_i^T \beta + \bar{\pi}/(1 - \bar{\pi}) (\mu_i - b_{jk} B_j(\beta_i) B_k(\eta_i))) \tau_2 \sum B_j^2(x_i) B_k^2(\eta_i) \right] \sigma_b^2$, and variance parameter $1/\sigma_b^2 = \tau_1 \sum B_j^0(x_i) B_k^1(\eta_i) z_i \tau_2 \sum B_j^2(x_i) B_k^2(\eta_i) (1 - z_i) \bar{\pi}^2 / (1 - \bar{\pi})^2 + 1$. Since the posterior distribution for $\bar{\pi}$ is not a standard distribution, we can apply the Metropolis Hastings algorithm for drawing posterior samples.
3.4 Simulation Studies

We consider two simulation settings. In the simulations, we assume \( \mu(x, \eta) = \sum_{j,k} b_{jk} B_j^p(x_i) B_k^1(\eta_i) \) where \( B_j^p \) are chosen to be radial basis functions. Let \( q_1 \) and \( q_2 \) be the number of basis functions \( B_j^p(x) \) and \( B_k^1(\eta) \). Suppose the knots for the basis functions are fixed. For \( i = 1, \ldots, n \), we generate
\[
b_{jk} \sim N(0, \psi_j \tau_k), j = 1, \ldots, q_1, \text{and } k = 1, \ldots, q_2.
\]
\[
x_i \sim N(0, 1), i = 1, \ldots, n
\]
\[
\eta_i \sim U(0, 1), i = 1, \ldots, n
\]
\[
\mu(x_i, \eta_i) = \sum_{j,k} b_{jk} B_j^p(x_i) B_k^1(\eta_i),
\]
\[
\epsilon_i \sim \tilde{\pi} N(\mu(x_i, \eta_i), \sigma_1^2) + (1 - \tilde{\pi}) N\left(\frac{-\pi \mu(x_i, \eta_i)}{(1 - \tilde{\pi})}, \sigma_2^2\right),
\]
\[
y_i = x_i^T \beta^* + \epsilon_i, i = 1, \ldots, n.
\]

We set \( \tilde{\pi} = 0.9 \), \( q_1 = 4 \), \( q_2 = 5 \), \( \sigma_1^2 = 0.1 \) and \( \sigma_2^2 = 0.4 \). For the first setting, we let \( p = 1 \) and the true \( \beta^* \) to be 2, whereas for the second setting, we let \( p = 3 \) and the true \( \beta^* \) to be \((-3, 2, 5)\).

We compare our model with Gaussian model and the homoscedastic skewed double exponential model. To compare the results, we use mean square error for the parameter \( \beta \), \( MSE = \sum_{m=1}^{M} (\hat{\beta}_m - \beta^*)^2 / M \), where \( \hat{\beta}_m \) is the posterior sample mean obtained in the \( m \)-th simulated data set. We also compare the prediction error mean for test data \( \sum_{m=1}^{M} \sum_{i=1}^{n_{test}} (y_i - x_i^T \hat{\beta}_m)^2 / M \) where \( \hat{\beta} = (1/M) \sum_{m=1}^{M} \hat{\beta}_m \). For each of the simulation settings (Table 3.1 and Table 3.2), we consider two sets of sample sizes - (50, 100) for training data and (50, 100) for test data. For Table 3.1, we generate 250 replicated dataset with \( \kappaappa = 2 \) for both latent variable and the covariate \( x \). For Table 3.2, we generate 200 replicated dataset with \( \kappaappa = 1 \) for the latent variable and \( \kappaappa = (0.2, 0.3, 0.1) \) for the corresponding dimensions of the covariate \( x \). We can see that our model performs much better than the Gaussian model and Skewed Double Exponential model comparing the MSE of coefficient parameters for both simulation studies.

3.5 An Epidemiological Application

In this section, we apply our model to a real data set from epidemiology studies. In epidemiology studies, a common focus is on assessing changes in a response distribution with a continuous exposure, adjusting for covariates. For example, Longnecker et al. (2001) studied the association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. DDT is highly effective against most malaria-transmitting mosquitoes and is being widely used in malaria-endemic areas. The Longnecker et al. (2001) study based on
Table 3.1: Approximate (via Monte Carlo) sampling mean and mean square error ($1000 \times \text{MSE}$ within parenthesis) of the estimated regression parameters ($\beta$); True parameter values are $\beta = 2$.

<table>
<thead>
<tr>
<th>Models</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 50$</td>
<td></td>
</tr>
<tr>
<td>Gaussian</td>
<td>1.9815(4.6)</td>
</tr>
<tr>
<td>SDE</td>
<td>2.0241(4.3)</td>
</tr>
<tr>
<td>LV-RD</td>
<td>1.9886(3.5)</td>
</tr>
<tr>
<td>$n = 100$</td>
<td></td>
</tr>
<tr>
<td>Gaussian</td>
<td>1.9846(2.6)</td>
</tr>
<tr>
<td>SDE</td>
<td>1.9908(2.5)</td>
</tr>
<tr>
<td>LV-RD</td>
<td>1.9904(1.8)</td>
</tr>
</tbody>
</table>

Table 3.2: Approximate (via Monte Carlo) sampling mean and mean square error ($1000 \times \text{MSE}$ within parenthesis) of the estimated regression parameters ($\beta_1, \beta_2, \beta_3$); True parameter values are $\beta_1 = -3, \beta_2 = 2, \beta_3 = 5$.

<table>
<thead>
<tr>
<th>Models</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 50$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaussian</td>
<td>-2.9897 (4.0)</td>
<td>1.9963 (4.4)</td>
<td>4.9693 (6.1)</td>
</tr>
<tr>
<td>SDE</td>
<td>-2.9672 (4.8)</td>
<td>1.9763 (4.1)</td>
<td>5.0045 (3.3)</td>
</tr>
<tr>
<td>LV-RD</td>
<td>-2.9824 (3.8)</td>
<td>1.9900 (3.8)</td>
<td>4.9695 (5.0)</td>
</tr>
<tr>
<td>$n = 100$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaussian</td>
<td>-2.9958 (2.1)</td>
<td>1.9949 (1.8)</td>
<td>4.9922 (2.6)</td>
</tr>
<tr>
<td>SDE</td>
<td>-2.9710 (2.9)</td>
<td>1.9783 (2.0)</td>
<td>5.0196 (2.1)</td>
</tr>
<tr>
<td>LV-RD</td>
<td>-2.9892 (2.0)</td>
<td>1.9891 (1.6)</td>
<td>4.9890 (2.2)</td>
</tr>
</tbody>
</table>

the US Collaborative Perinatal Project (CPP). It measured DDE concentration in mother’s serum samples stored during the third trimester of pregnancy, while also recording the gestational age at delivery, GAD, and other demographic factors, such as age. Complete data were available for 2380 children, of whom 361 were born preterm and 221 were small-for-gestational age.

Following standard practice in reproductive epidemiology, Longnecker et al. (2001) dichotomized GAD using a 37-week cut-off, so that deliveries occurring prior to 37 weeks of completed gestation were classified as preterm. Categorizing DDE using quantiles based on the empirical distribution, they fitted a logistic regression model, reporting evidence of a highly significant dose-response trend. Premature deliveries occurring earlier in the period before 37 weeks have greater risk of mortality and morbidity. Hence, from a public health and clinical perspective, it is of interest to assess how the entire left tail of the GAD distribution changes with DDE does, with effects earlier in gestation more important (Dunson and Park, 2008). Dunson
and Park (2008) analyzed the data by modeling the conditional density of $y \mid x$ using kernel stick-breaking processes. For ease of interpretation of covariate effects while still having the residual density changing flexibly with the predictors, they use model 3.2 to fit the dataset. There are 2313 children in the study, excluding the children for whom GAD exceeded 45 weeks, unrealistically high values attributable to measurement error. In this paper, we analyze the data for these 2313 children. The response variable we are interested in is the logarithm of gestational age at delivery. The predictor is DDE and the intercept term. The following histogram in Figure 3.2 shows the shape of the distribution of response.

![Histogram of log(GAD)](image)

Figure 3.2: Histogram of GAD

We analyzed the data using Gaussian model, Skewed Double Exponential (SDE) model and our LV-RD model. Figure 3.3 plots logarithm of the ratio of the Condi-
tional Predictive Ordinate (CPO) of the LV-RD model and CPO of Gaussian model against the standardized DDE, and the logarithm of the ratio of the CPO of the LV-RD model and CPO of SDE model against the standardized DDE. A value greater than 0 for this supports a LV-RD model over a Gaussian model or SDE model. In this data example, 67.92% (1571/2313) of observations favor the LV-RD model over the Gaussian model, i.e., a substantially higher proportion of observations supporting the LV-RD model over Gaussian model. Meanwhile, 66.58% (1540/2313) of observations favor the LV-RD model over the SDE model.

Figure 3.3: Plot of the log-ratio of CPOs obtained from LV-RD and Gaussian, from LV-RD and SDE model (y-axis), versus DDE (x-axis).

Figure 3.4, Figure 3.5 and Figure 3.6 show the the heteroscedasticity of observed response to the standardized predictor DDE, along with the first, second and third quantiles of the predicted values of logarithm of gestational age at delivery conditional on DDE (we apply 100 grids for DDE. For each grid, we draw 2000 posterior samples and calculate all the quantiles). We can see that our LV-RD model show the heteroscedasticity.

3.6 Discussion

In this paper, we created a new class of probability density functions which can handle the skewed and heteroscedastic error. Our density function can cover a large class of functions. The parameter $\tilde{\pi}$ can adjust the skewness and heteroscedasticity. Our proposed latent variable residual density performs much better than the Gaussian model and the skewed double exponential model comparing the mean square error in the simulation studies. The $\kappa$ in the Gaussian process $\mu(x, \eta)$ takes a critical important role. When $\kappa$ is large, $\mu(x, \eta)$ will concentrate at the point 0, we will get a density function, which mixture of two normal distributions with both modes at
Figure 3.4: Predicted quantiles versus DDE: Gaussian model

0. This way, the advantage of our model will not be apparent. However, when the $\mu(x, \eta)$ is large enough, the error distribution will be skewed and heteroscedastic, our model will have much better performance than the other two models. Meanwhile, the value of $\mu(x, \eta)$ also depends on $\|t - t_j\|^2$, thus, how to choose the optimal value of $\kappa$ based on the covariate matrix will be an interesting topic for the future work. A prior for $\kappa$ or some grid search methods may be applied.
Figure 3.5: Predicted quantiles versus DDE: SDE model
Figure 3.6: Predicted quantiles versus DDE: LV-RD model
APPENDIX A

THE PROOF OF THEOREMS IN CHAPTER 1

A.1 Proof of Equation 1.4

Since $f$ is decreasing, define a measure $\mu$ on $([0, \infty), B([0, \infty)))$ given by

$$
\mu([x, \infty)) = f(x).
$$

Observe that,

$$
f(x) = \mu([x, \infty)) = \int_{[0, \infty)} I_{[x, \infty)}(y) \mu(dy)
= \int_{[0, \infty)} \frac{1}{y} I_{[0,y]}(x) \{y \mu(dy)\}
$$

Set $G(dy) = y \mu(dy)$. The proof is completed if we can show that $G$ is a probability measure on $([0, \infty), B([0, \infty)))$.

$$
\int_{[0, \infty)} y \mu(dy) = \int_{y=0}^{\infty} \left\{ \int_{x=0}^{y} dx \right\} \mu(dy) = \int_{x=0}^{\infty} \left\{ \int_{y=x}^{\infty} \mu(dy) \right\} dx \ (\text{by Fubini})
= \int_{x=0}^{\infty} f(x) dx = 1.
$$

Hence $f(x) = \int_{[0,\infty)} \frac{1}{y} I_{[0,y]}(x) G(dy)$.

A.2 Proof of correlation

Let $F(\epsilon) = \Phi(\epsilon^*)$, then $\epsilon^* = \Phi^{-1}(F(\epsilon)) = h(\epsilon)$. From the Delta Method:

$$
\text{Var}(h_r) = \sum_i \left( \frac{\partial h_r}{\partial \epsilon_i} \right)^2 \text{Var}(\epsilon_i) + \sum_i \sum_{j \neq i} \left( \frac{\partial h_r}{\partial \epsilon_i} \right) \left( \frac{\partial h_r}{\partial \epsilon_j} \right) \text{Cov}(\epsilon_i, \epsilon_j)
$$

$$
\text{Cov}(h_r, h_s) = \sum_i \left( \frac{\partial h_r}{\partial \epsilon_i} \right) \left( \frac{\partial h_s}{\partial \epsilon_i} \right) \text{Var}(\epsilon_i) + \sum_i \sum_{j \neq i} \left( \frac{\partial h_r}{\partial \epsilon_i} \right) \left( \frac{\partial h_s}{\partial \epsilon_j} \right) \text{Cov}(\epsilon_i, \epsilon_j)
$$

58
For our model, $\frac{\partial h_r}{\partial \epsilon_i} = 0$ if $r \neq i$. So, we have:

\[
Var(h_r) = (\frac{\partial h_r}{\partial \epsilon_r})^2 Var(\epsilon_r)
\]

\[
Cov(h_r, h_s) = (\frac{\partial h_r}{\partial \epsilon_r})(\frac{\partial h_s}{\partial \epsilon_s}) Cov(\epsilon_r, \epsilon_s)
\]

Then, the correlation between $h_r$ and $h_s$ is approximately the same as the correlation between $\epsilon_r$ and $\epsilon_s$. 
APPENDIX B

PROOF FOR CHAPTER 3

B.1 prior for $\gamma$

As $g_{\eta}(y)$ is monotone transformation, this will lead $P(g_{\eta}(y_i) < g_{\eta}(x_i\beta)) = 0.5$, which means $P(\gamma_i + \epsilon_i < 0) = 0.5$. Suppose $\gamma_i$ has a prior density $f_{\gamma}(\gamma)$, then

$$P(\gamma_i + \epsilon_i < 0) = \int_{-\infty}^{\infty} \Phi(\epsilon) f_{\gamma}(u) du$$

$$= \int_{0}^{\infty} \Phi(\epsilon) f_{\gamma}(u) du + \int_{-\infty}^{0} \Phi(\epsilon) f_{\gamma}(u) du$$

$$= \int_{0}^{\infty} [1 - \Phi(\epsilon)] f_{\gamma}(u) du + \int_{0}^{\infty} \Phi(\epsilon) f_{\gamma}(-u) du$$

$$= \int_{0}^{\infty} f_{\gamma}(u) du + \int_{0}^{\infty} [f_{\gamma}(-u) - f_{\gamma}(u)] \Phi(\epsilon) du$$

If $f_{\gamma}(u) = f_{\gamma}(-u)$, then $P(\gamma_i + \epsilon_i < 0) = 0.5$. It is a sufficient condition.
APPENDIX C

THE PROOF OF THEOREMS IN CHAPTER 2

C.1 Proof of Proposition 1

1. 
\[
\int_{\mathbb{R}} f_0(\epsilon | x) = \int_{\mathbb{R}} \bar{\pi} f_0^\star(\epsilon | x) + \frac{(1 - \bar{\pi})^2}{\bar{\pi}} f_0\left(\frac{- (1 - \bar{\pi})}{\bar{\pi}} \epsilon | x\right)d\epsilon,
\]
\[
= \bar{\pi} + \frac{(1 - \bar{\pi})^2}{\bar{\pi}} \int_{\mathbb{R}} f_0^\star(\epsilon | x) - \frac{\bar{\pi}}{1 - \bar{\pi}} d\bar{\epsilon},
\]
\[
= \bar{\pi} + \frac{(1 - \bar{\pi})^2}{\bar{\pi}} \frac{\bar{\pi}}{1 - \bar{\pi}},
\]
\[
= 1
\]

Also,
\[
\int_{\mathbb{R}} \epsilon f_0(\epsilon | x) = \int_{\mathbb{R}} \bar{\pi} \epsilon f_0^\star(\epsilon | x) + \frac{(1 - \bar{\pi})^2}{\bar{\pi}} \epsilon f_0\left(\frac{- (1 - \bar{\pi})}{\bar{\pi}} \epsilon | x\right)d\epsilon,
\]
\[
= \bar{\pi} \mu_0^\star + \frac{(1 - \bar{\pi})^2}{\bar{\pi}} \left(\frac{- \bar{\pi}}{1 - \bar{\pi}}\right)^2 \int_{-\infty}^{\infty} \bar{\epsilon} f_0^\star(\bar{\epsilon} | x)d\bar{\epsilon},
\]
\[
= \bar{\pi} \mu_0^\star - \frac{(1 - \bar{\pi})^2}{\bar{\pi}} \left(\frac{- \bar{\pi}}{1 - \bar{\pi}}\right)^2 \mu_0^\star,
\]
\[
= 0
\]

Hence $\mathcal{F}_{t_0} \subset \mathcal{F}_r$.

2. Let $\mu_0^*(t, x) = (F_0^\star)^{-1}(t | x)$ where $F_0^\star$ is the cdf of $f_0^\star$. Since $\int_{[0,1]^p} \int_0^1 |\mu_0^*(t, x)||q(x)dtdx = \int_{[0,1]^p} \int_\mathbb{R} |y| f_0^\star(y | x)q(x)dydx < \infty$, $\mu_0^*q \in L_1[0,1]^{p+1}$. Fix $\epsilon > 0$. Since $\mu_0(\cdot | x) \notin C[0,1]$, the main idea is to find a continuous function $\tilde{\mu}_0^*$ close to $\mu_0^*$ in $L_1$ norm and exploit the fact that the prior on $\mu$ places positive mass to arbitrary sup-norm neighborhoods of $\tilde{\mu}_0^*$. The details are provided below.
Observe that \(||\phi_\sigma * f_0(\cdot | x) - f_0(\cdot | x)||_1 \to 0\) pointwise in \(x\) as \(\sigma \to 0\). Also, \(||\phi_\sigma * f_0(\cdot | x) - f_0(\cdot | x)||_1 \leq 2\). By the dominated convergence theorem, \(\int_{x \in [0,1]} \phi_\sigma * f_0(\cdot | x) - f_0(\cdot | x) ||q(x)|| dx \to 0\) as \(\sigma \to 0\). Find \(\delta > 0\) such that \(\int_{x \in [0,1]} \phi_\sigma * f_0(\cdot | x) - f_0(\cdot | x) ||q(x)|| dx < \epsilon/2\) for \(\sigma < \sigma_1\). Pick any \(\sigma_0 < \sigma_1\). Since \(C[0,1]^{p+1}\) is dense in \(L_1[0,1]^{p+1}\), for any \(\delta > 0\), we can find a \((x,y) \to \tilde{\mu}_0(x,y)\) continuous function \(\tilde{\mu}_0\) such that \(||\mu - \tilde{\mu}_0||_\infty < \delta\). Now, \(\int ||f_{\mu,\sigma,\tilde{\pi}}(\cdot | x) - f_{\tilde{\mu}_0,\sigma,\tilde{\pi}_0}(\cdot | x)||_1 q(x) dx \leq C \int ||\mu(\cdot | x) - \tilde{\mu}_0(\cdot | x)||_1 q(x) dx / \sigma\) for a global constant \(C\). Thus, for \(\delta = \epsilon \sigma_0/4\) and \(\kappa\) suitably small

\[
\left\{ f_{\mu,\sigma,\tilde{\pi}} : \sigma_0 < \sigma < \sigma_1, |\tilde{\pi} - \tilde{\pi}_0| < \kappa, ||\mu - \tilde{\mu}_0||_\infty < \delta \right\} \subset \left\{ f_{\mu,\sigma,\tilde{\pi}} : \int ||f_0 - f_{\mu,\sigma,\tilde{\pi}}||_1 q(x) dx < \epsilon \right\},
\]

since

\[
||f_0(\cdot | x) - f_{\mu,\sigma,\tilde{\pi}}(\cdot | x)||_1 < ||f_0(\cdot | x) - f_{\tilde{\mu}_0,\sigma}(\cdot | x)||_1 + ||f_{\tilde{\mu}_0,\sigma}(\cdot | x) - f_{\mu,\sigma,\tilde{\pi}(\cdot | x)||}_1.
\]

Thus, \(\Pi(g \in \mathcal{F}_0 : ||f_0 - g||_{1,q} < \epsilon) > \Pi_\mu(||\mu - \tilde{\mu}_0||_\infty < \delta) \Pi_\sigma(\sigma_0 < \sigma < \sigma_1) > 0\), since \(\Pi_\mu\) has full sup-norm support and \(\Pi_\sigma\) has full support on \([0,\infty)\).
REFERENCES


Yuanyuan Tang was born in 1984, Huainan, China. In the fall of 2002, she attended University of Science and Technology of China and completed her Bachelor’s degree in Mathematics and Applied Mathematics in the summer of 2006. She studied Applied/Computational Mathematics in Florida State University and obtained a Master of Science degree in the spring of 2009. In the fall of 2009, she was admitted to the doctoral program of Statistics at the Florida State University. She defended her dissertation in the summer of 2013.