Spring 2023

Advancements in Parametric Modal Regression

Qingyang Liu

Advancements in Parametric Modal Regression

by

Qingyang Liu

Bachelor of Economics
Hefei University 2013

Bachelor of Science
Northern Arizona University 2013

Master of Science
Temple Univeristy 2017

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2023

Accepted by:

Xianzheng Huang, Major Professor

Lianming Wang, Committee Member

Ray Bai, Committee Member

Bo Cai, Committee Member

Cheryl L. Addy, Interim Vice Provost and Dean of the Graduate School

Acknowledgments

I would like to express my sincere gratitude to the many individuals who have supported and guided me throughout my journey in statistics.

First and foremost, I extend my thanks to the professors at Hefei University and Northern Arizona University, who introduced me to the field of statistics at the undergraduate level. Without their guidance, I would not have had the opportunity to pursue statistics at the master's level.

I am also deeply grateful to Drs. William W.S. Wei, Chengyong Tang, and Zhigen Zhao, who taught me graduate-level courses such as mathematical statistics, linear regression models, and time series models. Their support and mentorship were invaluable to my success at Temple University and in securing my position as a Statistician at DuPont-Corteva Agriscience. I also extend my heartfelt appreciation to my former supervisor, Dr. Hua Mo, for providing me with challenging projects and opportunities to grow as a statistician.

I am fortunate to have completed my PhD at the Department of Statistics at the University of South Carolina, where I had the opportunity to work alongside many outstanding professors. Particularly, I want to thank Dr. Ray Bai, who served as a member of my PhD dissertation committee and is one of my collaborators. Dr. Ray Bai's guidance and high-quality instruction were instrumental in my research accomplishments. I am also grateful to Dr. Joshua Tebbs for his valuable advice on my academic career.

I would like to express my deepest appreciation to my wife, who is also a PhD candidate at the Biostatistics Department at the University of South Carolina. Her

unwavering support and encouragement have been vital to my success throughout my doctoral journey. I would also like to express my sincere gratitude to my parents, grandparents, and parents-in-law for their steadfast support.

Lastly, I want to express my profound gratitude to my PhD advisor, Dr. Xianzheng Huang, for her exceptional guidance and mentorship throughout my doctoral studies. Dr. Xianzheng Huang is an outstanding advisor and an inspiration to any student. Without her support, I would not have been able to achieve my academic and research accomplishments.

Thank you to everyone who has played a role in my academic and professional journey.

Abstract

This dissertation considers statistical inference methods for parametric modal regression models. In Chapter 1, we motivate the mode as the measure of central tendency instead of the median or the mean with an example. Following the motivational example, we include an overview of existing modal regression models. Later, in the same chapter, we explain advantages of the parametric modal regression models over existing nonparametric modal regression models. In Chapter 2, we address issues in statistical inference brought in by data contaminated with measurement error. With measurement error in covariates, statistical inference methods designed for modal regression models with error-free covariates become inappropriate. We use an innovative Monte-Carlo based method to revise the original log-likelihood function that one uses in the absence of covariates measurement error. This revision leads to a new objective function adequately accounting for measurement error that one maximizes with respect to unknown parameters in the regression model. We also propose a model diagnostic method based on parametric bootstrap for the parametric modal regression with error in covariates. The proposed method for estimating regression parameters is applicable for any parametric modal regression models. However, there are only a handful of existing distributions that are suitable for the modal regression model for heavy-tailed response data. To allow for flexible modal regression, we propose a new unimodal distribution called flexible Gumbel distribution in Chapter 3. We present both frequentist and Bayesian inference methods for the flexible Gumbel distribution in the same chapter. Chapter 4 introduces the general unimodal distribution family that encompasses a range of unimodal asymmetric distributions and incorporates the

flexible Gumbel distribution as a specific instance. Based on the general unimodal distribution family, we propose a unified framework for Bayesian modal regression that is well-suited for analyzing asymmetric and fat-tailed data. We propose the Gaussian process modal regression model in Chapter 5. Unlike the classic Gaussian process regression model where one assumes a Gaussian process for the conditional mean of the response, in our proposed Gaussian process regression model, we assume a Gaussian process for the conditional mode.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Three Measures of Central Tendency

The mean, median, and mode are the three most commonly used measures of central tendency of data. For data from heavy tailed and skewed distribution, the mode is a more sensible measure of central tendency than the mean or median. Because of the ubiquity of heavy-tailed and skewed data in biology, sociology, economics, and many other fields, there are plenty examples where the mode is a more appropriate measure of central location than the other two (Chacón, 2020). An example about the highly right-skewed distribution of deposits from saving institutions in the United States is provided in Section 1.2.

1.2 An Example: Deposits of Large Banks and Savings Institutions

To economists, it is no secret that wealth distributions are highly skewed to the right (Benhabib and Bisin, 2018). The cumulative nature of wealth has impact on not only wealth status of individuals, but also the deposits of bank holding companies. The dataset with records of deposits of 50 banks and savings institutions is collected from Siegel (2012, Table 3.4.1). Figure 1.1 presents a kernel density estimate for the deposits (in billions of dollars) of large banks and savings institutions in the United States on July 2, 2010. The sample mean, which equals 92.6 billion dollars, is obviously not a good measure of central tendency for most large banks and savings institutions in the United States. In fact, 40 of 50 banks and savings institutions in

the dataset have deposit less than 92.6 billion dollars. In the meanwhile, the sample median is 40.5 billion dollars. Since the sample median is resistant to outliers, the large difference between sample median and sample mean should not be surprising. However, the location of the sample median is not self explained. In other words, it is difficult to "guess" the location of the sample median based solely on the estimated density plot. The sample mode, which locates right under the peak of density curve by its definition, can be easily interpreted as that banks are most likely to have deposits of 28.4 billion dollars.

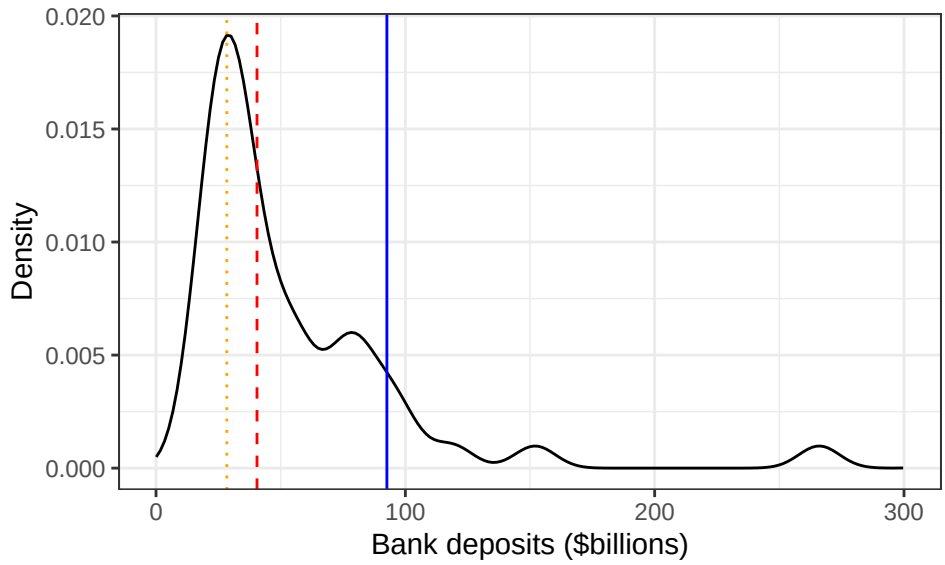


Figure 1.1 Density estimation for the the deposits (in billions of dollars) of large banks and savings institutions in the United States on July 2, 2010. From left to right, the orange dotted line, red dashed line, and blue solid line represent the estimated mode, median, and mean, respectively.

1.3 Overview of Existing Modal Regression Models

While there exists an extensive literature on regression models that relate the mean or median of a response variable Y to covariates $\mathbf{X}$, there are much less work on regression models tailored for the conditional mode of Y given $\mathbf{X}$ (Sager and Thisted, 1982; Lee,

1989, 1993). Among the limited existing modal regression methods, the majority of them are in the semi-/non-parametric framework (Yao and Li, 2013; Chen et al., 2016), which typically suffer low statistical efficiency when comparing with their parametric counterparts. One reality that discourages use of parametric models for inferring the mode is that very few named distributions that allow asymmetry can be conveniently formulated as distribution families indexed by the mode along with other parameters. Among the few groups of authors who considered parametric modal regression models, Aristodemou (2014, Chapter 3) assumed a gamma distribution for a non-negative response with a covariate-dependent mode; Bourguignon et al. (2020) followed a similar model construction while also allowing a covariate-dependent precision parameter for the gamma distribution. Focusing on bounded response data, Zhou and Huang (2020) proposed two modal regression models, one based on a beta distribution and the other based on a generalized biparabolic distribution for the response given covariates. In all three aforementioned works, frequentist likelihood-based methods are developed to infer model parameters. Most recently, Zhou and Huang (2022) unified the mean regression and modal regression in a Bayesian framework by reparameterizing a four-parameter beta distribution with an unknown support so that the mean or the mode of Y depends on $\mathbf{X}$. Earlier works on Bayesian modal regression, including parametric and nonparametric methods, can also be found in Aristodemou (2014, Chapter 2).

Chapter 2

Parametric Modal Regression with Error in Covariates

2.1 Introduction

All works on modal regression cited in chapter 1 assume that covariates are measured precisely. Data analysts in many disciplines are well aware that, among all variables of interest, some of them often cannot be measured precisely due to inaccurate measuring devices or human error in data collection. Some variables are in principle inaccessible and only some surrogates of them can be measured. For example, one's long-term blood pressure is an important biomarker associated with one's heart health, yet it cannot be directly measured. Instead, measurable surrogates of it are blood pressure readings collected during a doctor's visit, which can be viewed as error-contaminated versions of one's long-term blood pressure. It has also been well-understood that ignoring covariates measurement error in mean regression or quantile regression usually lead to misleading inference results. There exists a large collection of works on mean regression methodology accounting for measurement error (Carroll et al., 2006; Fuller, 2009; Buonaccorsi, 2010; Yi, 2017), and also some works in quantile regression to address this complication (He and Liang, 2000; Wei and Carroll, 2009; Wang et al., 2012). Modal regression methodology that address this issue only emerged recently, including those developed by Zhou and Huang (2016), Li and Huang (2019), and Shi et al. (2021), all of which opted for a nonparametric model for the error term in the primary regression model. There is a lack of methodology to account for error-prone

covariates in parametric modal regression, and our study presented in this chapter fills the void.

In preparation for proposing a method to account for measurement error in covariates that is applicable to any parametric modal regression models, we first formulate the measurement error model and discuss complications unique to modal regression models in Section 2.2. For concreteness, we then focus on the beta modal regression model for a response supported on [0, 1] with an error-prone covariate, and propose consistent estimation methods to infer model parameters that account for measurement error in Section 2.3. A model diagnostic method is developed to detect model misspecifications when adopting the beta modal regression model in a given application in Section 2.4. Simulation studies are reported in Section 2.5 to demonstrate the performance of the estimation and diagnostics methods. We apply the proposed modal regression method accounting for covariate measurement error to data sets arising from two real-life studies in Section 2.6, where we also discuss revisions of the method to adapt to more general settings. Section 2.7 gives concluding remarks and future research directions.

2.2 Data and Model

2.2.1 Observed data

Suppose that, given p covariates in $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$, Y follows a unimodal distribution specified by the probability density function (pdf), $f_{Y|\mathbf{X}}(y|\mathbf{x})$. Denote by $\theta(\mathbf{x})$ the mode of Y given $\mathbf{X} = \mathbf{x}$. In modal regression without measurement error, one infers $\theta(\mathbf{x})$ based on a random sample of size n from the joint distribution of $(Y, \mathbf{X})$, $\{(Y_j, \mathbf{X}_j)\}_{j=1}^n$, where $\mathbf{X}_j = (X_{1,j}, \ldots, X_{p,j})^{\mathrm{T}}$. Now suppose that a covariate in $\mathbf{X}$, say, X_1, is prone to measurement error, and a surrogate W is observed instead of X_1, with n_j replicate measures of $X_{1,j}$ in $\widetilde{W}_j = \{W_{j,k}\}_{k=1}^{n_j}$, for $j = 1, \ldots, n$. In this study, we

assume that $W_{j,k}$ relates to $X_{1,j}$ via an additive measurement error model,

$$W_{j,k} = X_{1,j} + U_{j,k}, \text{ for } j = 1, \ldots, n \text{ and } k = 1, \ldots, n_j, \tag{2.2.1}$$

where $\{U_{j,k},\, k = 1, \ldots, n_j\}_{j=1}^{n}$ are independent and identically distributed (i.i.d.) mean-zero measurement error, which are independent of $\{(Y_j, \mathbf{X}_j)\}_{j=1}^{n}$ to guarantee nondifferential measurement error as considered in the classical measurement error models (Carroll et al., 2006, Section 2.5).

In a naive univariate modal regression analysis using the surrogate data, one treats W as if it were $X = X_1$, and equivalently, views the conditional pdf of Y given $W = w$, $f_{Y|W}(y|w)$, the same as $f_{Y|X}(y|w)$. As a result, naive modal regression analysis essentially infers the mode of $f_{Y|W}(y|w)$ instead of $\theta(\cdot)$. In the context of univariate mean regression models not limited to linear regression, the attenuation effect of measurement error on covariate effect estimation is often noted in the literature (Carroll et al., 2006; Buonaccorsi, 2010), which causes the estimated covariate effect of a truly influential covariate to be pulled towards zero. Naive modal regression can suffer the same attenuation effect. For instance, if the mean and the mode of $f_{Y|X}(y|x)$ differ by a quantity that does not depend on covariates, such as for a Gumbel distribution that depends on a covariate X only via the mode but not via the scale parameter, then the impact of measurement error on naive inference for the conditional mean mostly carries over to naive inference for $\theta(x)$. In other model settings where the conditional mean and mode of Y differ by a quantity that does depend on the error-prone covariate, the effect of measurement error on naive modal regression demands investigation on a case-by-case basis. Even before conducting such investigation, a more fundamental question needs to be addressed, that is whether or not naive modal regression is meaningful, since unimodality of $f_{Y|X}(y|x)$ does not guarantee unimodality of $f_{Y|W}(y|w)$. Indeed, there is an extra layer of complication in modal regression with an error-prone covariate that does not exist in mean regression since, if the mean of Y given X, $\mu(X)$, is well defined, then the mean of Y given

W is $E\{\mu(X)|W\}$, which is also well defined in most settings of practical interest. Because of this additional complication, correcting naive inference to account for measurement error in modal regression is more challenging than the counterpart task in mean regression. For example, a strategy that can be easy to implement in mean regression is to correct the bias in a naive estimator of a parameter to produce an improved estimator accounting for measurement error (Carroll et al., 2006, Section 3.4). This idea of de-biasing naive estimation may not be a sensible approach now with the existence of a naive mode function in question.

2.2.2 Regression model

We propose to account for measurement error when inferring parameters in a modal regression model by exploiting the idea of corrected scores. In particular, we focus on modeling a bounded response Y, which is commonly encountered in practice, such as test scores, disease prevalence, and the fraction of household income spent on food. Any bounded response with a known support can be scaled to be supported on the unit interval [0, 1]. Beta distribution is a parametric family that encompasses various shapes of distributions supported on [0, 1], and thus serves as a relatively flexible basis for building a regression model for such responses. For a random variable V that follows a beta distribution with shape parameters $\alpha_1, \alpha_2 > 0$, i.e., $V \sim \text{beta}(\alpha_1, \alpha_2)$, its density function is,

$$f(v; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} v^{\alpha_1 - 1}(1 - v)^{\alpha_2 - 1}, \text{ for } 0 < v < 1,$$

where $\Gamma(\cdot)$ is the Gamma function. When α_1, $\alpha_2 > 1$, this distribution has a unique mode given by $\theta = (\alpha_1 - 1)/(\alpha_1 + \alpha_2 - 2)$. To prepare for modal regression, we reparameterize the beta distribution by setting $\alpha_1 = 1 + m\theta$ and $\alpha_2 = 1 + m(1 - \theta)$, where $m > 0$ plays the role of a precision parameter, with a larger value of m leading to a smaller variance of the distribution. By construction, as long as the mode $\theta \in (0, 1)$

exists, which we assume throughout the study, we have $\alpha_1, \alpha_2 > 1$ following this parameterization.

With a beta distribution family indexed by (θ, m) formulated, a beta modal regression model follows by introducing covariates-dependent mode of Y, $\theta(\mathbf{X}) = g(\boldsymbol{\beta}^{\mathrm{T}}\tilde{\mathbf{X}})$, where $\tilde{\mathbf{X}} = (1, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ with β_0 being the intercept and $\beta_1, \ldots, \beta_p$ representing covariate effects associated with the p covariates in $\mathbf{X}$, and $g(\cdot)$ is a user-specified link function, such as logit, probit, log-log, and complementary log-log. Now a modal regression model for Y is fully specified by the following conditional distribution of Y given $\mathbf{X}$,

$$Y|\mathbf{X} \sim \text{beta}(1 + m\theta(\mathbf{X}),\, 1 + m\{1 - \theta(\mathbf{X})\}). \tag{2.2.2}$$

Combining (2.2.2) with (2.2.1) completes the specification of a modal regression model for a response Y supported on [0, 1] and covariates $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$, with X_1 subject to additive nondifferential measurement error. The focal point of inference lies in parameters involved in the primary regression model in (2.2.2), $\boldsymbol{\Omega} = (\boldsymbol{\beta}^{\mathrm{T}}, m)^{\mathrm{T}}$. Parameters appearing in (2.2.1) are of secondary interest but required to specify the measurement error distribution.

2.3 Parameter estimation

2.3.1 Maximum likelihood estimation

In the absence of measurement error, one may carry out maximum likelihood estimation of $\boldsymbol{\Omega}$ straightforwardly by solving the normal score equations for $\boldsymbol{\Omega}$. More specifically, the log-likelihood of error-free data, $\mathcal{D} = \{(Y_j, \mathbf{X}_j)\}_{j=1}^n$, is

$$\begin{aligned}
\ell(\boldsymbol{\Omega}; \mathcal{D}) &= \sum_{j=1}^{n} \ell(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j) \\
&= n \log \Gamma(2 + m) - \sum_{j=1}^{n} \log \left(\Gamma(1 + m\theta\left(\mathbf{X}_j\right))\Gamma(1 + m\left\{1 - \theta\left(\mathbf{X}_j\right)\right\})\right) \\
&\quad + m \sum_{j=1}^{n} \left[\theta\left(\mathbf{X}_j\right) \log Y_j + \left\{1 - \theta\left(\mathbf{X}_j\right)\right\} \log\left(1 - Y_j\right)\right].
\end{aligned} \tag{2.3.1}$$

Differentiating (2.3.1) with respect to $\boldsymbol{\Omega}$ leads to the score equations, $\sum_{j=1}^{n} \boldsymbol{\Psi}_0(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j) = \mathbf{0}$, where the score vector evaluated at the j-th data point, $\boldsymbol{\Psi}_0(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j)$, consists of the following scores, for $j = 1, \ldots, n$,

$$\frac{\partial \ell(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j)}{\partial \boldsymbol{\beta}} = \left\{ -m\psi(1 + m\theta(\mathbf{X}_j)) + m\psi(1 + m\{1 - \theta(\mathbf{X}_j)\}) + m \log\left(\frac{Y_j}{1 - Y_j}\right) \right\} \times g'(\boldsymbol{\beta}^{\mathrm{T}} \tilde{\mathbf{X}}_j) \tilde{\mathbf{X}}_j, \tag{2.3.2}$$

$$\frac{\partial \ell(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j)}{\partial m} = \psi(2 + m) - \theta(\mathbf{X}_j)\psi(1 + m\theta(\mathbf{X}_j)) - \{1 - \theta(\mathbf{X}_j)\}\psi(1 + m\{1 - \theta(\mathbf{X}_j)\}) + \theta(\mathbf{X}_j) \log Y_j + \{1 - \theta(\mathbf{X}_j)\} \log(1 - Y_j), \tag{2.3.3}$$

where $\psi(t) = (d/dt) \log \Gamma(t)$ is the digamma function and $g'(t) = (d/dt)g(t)$.

2.3.2 Monte-Carlo corrected scores

In the presence of measurement error, a naive estimator of $\boldsymbol{\Omega}$ solves the naive score equations resulting from replacing $X_{1,j}$ with $\overline{W}_j = n_j^{-1} \sum_{k=1}^{n_j} W_{j,k}$ in (2.3.2) and (2.3.3), for $j = 1, \ldots, n$. As pointed out earlier and also evidenced in simulation study to be presented later, this naive treatment typically results in misleading inference for $\boldsymbol{\Omega}$. We propose to follow the idea of the corrected score method (Nakamura, 1990) and revise the naive scores to obtain estimating equations that adequately account for measurement error. The thrust of the corrected score method is to use the observed error-prone data, $\mathcal{D}^* = \{(Y_j, \widetilde{W}_j, \mathbf{X}_{-1,j})\}_{j=1}^{n}$ with $\widetilde{W}_j = \{W_{j,k}\}_{k=1}^{n_j}$ and $\mathbf{X}_{-1,j} = (X_{2,j}, \ldots, X_{p,j})^{\mathrm{T}}$, to construct unbiased estimators of the above normal scores. In this vein of thinking, one treats $\{X_{1,j}\}_{j=1}^{n}$ as unknown parameters instead of realizations of a random variable, and thus one takes on the functional point of view as opposed to the structural viewpoint of measurement error models where a distribution for X_1 is assumed (Carroll et al., 2006, Section 2.1).

We begin with applying the Monte-Carlo-amenable method proposed by Stefanski et al. (2005), a method originating from the idea described in Stefanski (1989). More specifically, we construct a score, $\boldsymbol{\Psi}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \mathbf{X}_{-1,j})$, that satisfies

$E\{\boldsymbol{\Psi}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \mathbf{X}_{-1,j})|Y_j, \mathbf{X}_j\} = \boldsymbol{\Psi}_0(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j)$, for $j = 1, \ldots, n$. This particular method is especially suitable for settings with a univariate error-prone covariate subject to normal measurement error U. We will address violation of the normality assumption on U in Section 2.3, and describe revisions of the method to adapt to settings with multiple error-prone covariates in Section 2.6. As shown in Stefanski et al. (2005, Theorem 1), the minimum variance unbiased estimator of $\boldsymbol{\Psi}_0(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j)$ is given by

$$\boldsymbol{\Psi}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \mathbf{X}_{-1,j}) = E\left\{ \boldsymbol{\Psi}_0\left(\boldsymbol{\Omega}; Y_j, \overline{W}_j + i\sqrt{\frac{(n_j-1)S_j^2}{n_j}}T, \mathbf{X}_{-1,j}\right) \middle| Y_j, \overline{W}_j, S_j^2, \mathbf{X}_{-1,j} \right\}, \tag{2.3.4}$$

where i is the imaginary unit, S_j^2 is the sample variance of $\widetilde{W}_j = \{W_{j,k}\}_{k=1}^{n_j}$, and $T = Z_1/(\sum_{k=1}^{n_j-1} Z_k^2)^{1/2}$ is independent of all observed data, in which $Z_1, \ldots, Z_{n_j-1}$ are independent standard normal random variables. The estimator of $\boldsymbol{\Psi}_0(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j)$ in (2.3.4) originates from a jackknife exact-extrapolant estimator constructed for the purpose of estimating a function of the mean of a normal distribution based on a random sample from the distribution. In the context of (2.3.4), this random sample is $\widetilde{W}_j$ from $N(X_{1,j}, \sigma_u^2)$, where σ_u^2 is the measurement error variance, i.e., assuming $U \sim N(0, \sigma_u^2)$ in (2.2.1), and the function of the normal mean $X_{1,j}$ is $\boldsymbol{\Psi}_0(\boldsymbol{\Omega}; Y_j, X_{1,j}, \mathbf{X}_{-1,j})$. The expectation in (2.3.4) cannot be derived in closed form. But since the only quantity viewed as random when deriving this conditional expectation is T that is independent of observed data, one can estimate this expectation unbiasedly via an empirical mean based on simulated random samples of T. Moreover, as shown in Stefanski et al. (2005), even though (2.3.4) is complex-valued by construction, the expectation of its imaginary part is zero as long as $\boldsymbol{\Psi}_0(\boldsymbol{\Omega}; Y_j, X_{1,j}, \mathbf{X}_{-1,j})$ is infinitely differentiable with respect to $X_{1,j}$, which is guaranteed in our case by choosing a link function $g(t)$ that is infinitely differentiable. Hence, using the real part of the empirical version of (2.3.4) suffices for constructing an unbiased estimator of $\boldsymbol{\Psi}_0(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j)$. This leads to the following corrected score based on a simulated random sample of T

of size B, $\widetilde{T}_j = \{T_{j,b}\}_{b=1}^{B}$, for $j = 1, \ldots, n$,

$$\boldsymbol{\Psi}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j}) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{Re}\left\{ \boldsymbol{\Psi}_0\left(\boldsymbol{\Omega}; Y_j, \overline{W}_j + i\sqrt{\frac{(n_j-1)S_j^2}{n_j}} T_{j,b}, \mathbf{X}_{-1,j}\right)\right\}, \tag{2.3.5}$$

where $\mathrm{Re}(t)$ denotes the real part of a complex-valued t.

One now can solve the following system of $p+2$ equations based on the corrected score in (2.3.5),

$$\sum_{j=1}^{n} \boldsymbol{\Psi}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j}) = \mathbf{0}, \tag{2.3.6}$$

for $\boldsymbol{\Omega}$ to obtain a consistent estimator $\hat{\boldsymbol{\Omega}}$, where $\widetilde{T}_1, \ldots, \widetilde{T}_n$ are independent. Solving (2.3.6) for $\boldsymbol{\Omega}$ is equivalent to solving an optimization problem, that is,

$$\hat{\boldsymbol{\Omega}} = \underset{\boldsymbol{\Omega}\in\mathbb{R}^{p+1}\times\mathbb{R}^{+}}{\arg\min} \left\{\sum_{j=1}^{n} \boldsymbol{\Psi}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j})\right\}^{\mathrm{T}} \left\{\sum_{j=1}^{n} \boldsymbol{\Psi}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j})\right\}. \tag{2.3.7}$$

The equivalence between (2.3.7) and the solution to (2.3.6) is obvious when there exists a unique solution to (2.3.6). An added benefit of dealing with an optimization problem is more appreciated in the presence of model misspecification that can potentially lead to non-existence of a solution to (2.3.6), yet (2.3.7) may still be well-defined with meaningful statistical interpretations according to White (1982).

2.3.3 Monte-Carlo corrected log-likelihood

To this end, estimating $\boldsymbol{\Omega}$ appears to be a straightforward optimization problem. But the numerical procedure to obtain (2.3.7) requires evaluating $p+2$ scores at each iteration, which can be cumbersome and very demanding on the computer memory and central processing unit, especially due to the Monte Carlo nature of the score in (2.3.5) that involves computing a vector-valued score B times. Viewing the quadratic form in (2.3.7) as an objective function that accounts for measurement error, we propose to use a different objective function that also takes measurement error into account and is computationally less cumbersome to optimize. This new objective

function is obtained by correcting the naive log-likelihood function $\ell(\mathbf{\Omega}; Y_j, \overline{W}_j, \mathbf{X}_{-1,j})$ that is the summand of (2.3.1) with $X_{1,j}$ evaluated at $\overline{W}_j$, for $j = 1, \ldots, n$. Similar to the construction of the corrected score in (2.3.5) based on the naive score, the new objective function based on the naive log-likelihood evaluated at the j-th observed data point is

$$\tilde{\ell}(\mathbf{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j}) = \frac{1}{B} \sum_{b=1}^{B} \mathrm{Re} \left\{ \ell \left(\mathbf{\Omega}; Y_j, \overline{W}_j + i \sqrt{\frac{(n_j - 1) S_j^2}{n_j}} T_{j,b}, \mathbf{X}_{-1,j} \right) \right\}, \tag{2.3.8}$$

which satisfies $E\{\tilde{\ell}(\mathbf{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j}) | Y_j, \mathbf{X}_j\} = \ell(\mathbf{\Omega}; Y_j, \mathbf{X}_j)$, for $j = 1, \ldots, n$. We then define an estimator of $\mathbf{\Omega}$ as

$$\hat{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \in \mathbb{R}^{p+1} \times \mathbb{R}^+}{\arg\max} \sum_{j=1}^{n} \tilde{\ell}(\mathbf{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j}), \tag{2.3.9}$$

which only requires repeated evaluation of a scalar function in (2.3.8) at each iteration of an optimization algorithm. In simulation studies (not presented in this chapter) where we estimate $\mathbf{\Omega}$ using these two routes of optimization according to (2.3.7) and (2.3.9), we obtain very similar estimates of $\mathbf{\Omega}$, with the former route more computationally demanding than the latter. The numerical similarity of (2.3.7) and (2.3.9) may be expected given the connection between the naive score and the naive log-likelihood, in addition to the equivalence between the solution to the normal score equation and the maximum likelihood estimator in the absence of measurement error. We refer to the estimator defined in (2.3.9) the Monte Carlo corrected log-likelihood estimator, or MCCL for short.

Whether one follows the idea of correcting the naive scores or the route of correcting the naive log-likelihood to account for measurement error, our proposed estimation method falls in the general framework of M-estimation (Boos and Stefanski, 2013, Chapter 7). As an M-estimator, the MCCL estimator $\hat{\mathbf{\Omega}}$ is a consistent estimator of $\mathbf{\Omega}$ that is asymptotically normal under regularity conditions stated in, for example, Theorem 7.2 in Boos and Stefanski (2013). Moreover, motivated by its asymptotic

variance of the sandwich form (Boos and Stefanski, 2013, Section 7.2.1), the variance of $\hat{\boldsymbol{\Omega}}$ can be estimated by

$$\mathbf{V}(\mathcal{D}^*;\hat{\boldsymbol{\Omega}}) = \left\{\mathbf{A}(\mathcal{D}^*;\hat{\boldsymbol{\Omega}})\right\}^{-1}\mathbf{B}(\mathcal{D}^*;\hat{\boldsymbol{\Omega}})\left[\left\{\mathbf{A}(\mathcal{D}^*;\hat{\boldsymbol{\Omega}})\right\}^{-1}\right]^{\mathrm{T}}, \tag{2.3.10}$$

where

$$\begin{aligned}
\mathbf{A}(\mathcal{D}^*;\hat{\boldsymbol{\Omega}}) &= \frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial\boldsymbol{\Omega}^{\mathrm{T}}}\boldsymbol{\Psi}(\boldsymbol{\Omega};Y_j,\widetilde{W}_j,\widetilde{T}_j,\mathbf{X}_{-1,j})\bigg|_{\boldsymbol{\Omega}=\hat{\boldsymbol{\Omega}}},\\
\mathbf{B}(\mathcal{D}^*;\hat{\boldsymbol{\Omega}}) &= \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{\Psi}(\hat{\boldsymbol{\Omega}};Y_j,\widetilde{W}_j,\widetilde{T}_j,\mathbf{X}_{-1,j})\left\{\boldsymbol{\Psi}(\hat{\boldsymbol{\Omega}};Y_j,\widetilde{W}_j,\widetilde{T}_j,\mathbf{X}_{-1,j})\right\}^{\mathrm{T}}.
\end{aligned}$$

2.4 Model diagnostics

Even though we avoid specifying the true covariate distribution by adopting the functional viewpoint of measurement error models, the primary regression model in (2.2.2) is fully parametric. This raises the concern of model misspecification and calls for model diagnostics tools. Model diagnostics based on error-prone data is more challenging than settings without measurement error. In particular, conventional residual-based diagnostics methods that require evaluating an estimated regression function, whether it is the conditional mean $\mu(\mathbf{X})$ in mean regression or the conditional mode $\theta(\mathbf{X})$ in modal regression, are no longer applicable now that a true covariate is unobserved. Another contribution of our study is an effective score-based diagnostic tool that circumvents this obstacle a traditional residual-based diagnostic method faces in the presence of measurement error.

For the beta modal regression model without error in covariates, Zhou and Huang (2020) propose a score-based test statistic defined below for the purpose of model diagnostics,

$$Q(\hat{\boldsymbol{\Omega}}_0;\mathcal{D}) = \frac{n-2}{2(n-1)}\overline{\mathbf{S}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\overline{\mathbf{S}}, \tag{2.4.1}$$

where $\hat{\boldsymbol{\Omega}}_0$ is the maximum likelihood estimator of $\boldsymbol{\Omega}$, $\overline{\mathbf{S}} = n^{-1}\sum_{j=1}^{n}\mathbf{S}(\hat{\boldsymbol{\Omega}}_0;Y_j,\mathbf{X}_j)$, and $\hat{\boldsymbol{\Sigma}} = \{n(n-1)\}^{-1}\sum_{j=1}^{n}\{\mathbf{S}(\hat{\boldsymbol{\Omega}}_0;Y_j,\mathbf{X}_j)-\overline{\mathbf{S}}\}\{\mathbf{S}(\hat{\boldsymbol{\Omega}}_0;Y_j,\mathbf{X}_j)-\overline{\mathbf{S}}\}^{\mathrm{T}}$, in which, for

$j = 1, \ldots, n$,

$$\mathbf{S}(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j) = \begin{bmatrix} \log Y_j - \psi(1 + m\theta(\mathbf{X}_j)) + \psi(2 + m) \\ Y_j \log Y_j - \dfrac{\{1 + m\theta(\mathbf{X}_j)\}\{\psi(2 + m\theta(\mathbf{X}_j)) - \psi(3 + m)\}}{2 + m} \end{bmatrix} \tag{2.4.2}$$

is the score vector constructed by matching $\log V$ and $V \log V$ with their respective expectations for $V \sim \text{beta}(\alpha_1, \alpha_2)$, and thus $E\{\mathbf{S}(\boldsymbol{\Omega}; Y_j, \mathbf{X}_j)\} = \mathbf{0}$ in the absence of model misspecification. By construction, a larger value of the nonnegative $Q(\hat{\boldsymbol{\Omega}}_0; \mathcal{D})$ provides stronger evidence indicating model misspecification. A parametric bootstrap procedure is developed in Zhou and Huang (2020) to estimate the null distribution of $Q(\hat{\boldsymbol{\Omega}}_0; \mathcal{D})$, from which one onbtains an estimated p-value for the test.

Returning to our beta modal regression model with error-in-covariate, we apply the idea of corrected score here to construct a counterpart of (2.4.2) to obtain a score accounting for measurement error whose mean is zero in the absence of model misspecification. This yields the corrected score evaluated at the j-th observed data point for model diagnostics, for $j = 1, \ldots, n$,

$$\tilde{\mathbf{S}}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j}) = \frac{1}{B} \sum_{b=1}^{B} \text{Re} \left\{ \mathbf{S} \left(\boldsymbol{\Omega}; Y_j, \overline{W}_j + i \sqrt{\frac{(n_j - 1)S_j^2}{n_j}} T_{j,b}, \mathbf{X}_{-1,j} \right) \right\}. \tag{2.4.3}$$

The test statistic of the quadratic form denoted by $\tilde{Q}(\hat{\boldsymbol{\Omega}}; \mathcal{D}^*)$ that is parallel to (2.4.1) follows by using the MCCL estimator $\hat{\boldsymbol{\Omega}}$ instead of $\hat{\boldsymbol{\Omega}}_0$, replacing $\overline{\mathbf{S}}$ appearing in (2.4.1) with $n^{-1} \sum_{j=1}^{n} \tilde{\mathbf{S}}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j})$, and revising $\hat{\boldsymbol{\Sigma}}$ accordingly. But the next hurdle emerges, that is the design of a parametric bootstrap procedure for estimating the null distribution of $\tilde{Q}(\hat{\boldsymbol{\Omega}}; \mathcal{D}^*)$. Traditional parametric bootstrap in the regression setting, such as the procedure in Zhou and Huang (2020), involves generating response data from the primary regression model that again requires evaluating an estimated regression function at the true covariates that are partly unobserved in the current context. We overcome this hurdle by "estimating" unobserved true covariate data, as implemented in the method of regression calibration (Chapter 4, Carroll et al., 2006) that takes on the structural viewpoint of measurement error models.

Under the classical measurement error in (2.2.1), the best linear predictor of $X_{1,j}$ is $E(X_{1,j}|\overline{W}_j) = \mu_1 + \lambda_j(\overline{W}_j - \mu_1)$, where $\mu_1 = E(X_1)$ and $\lambda_j = n_j\sigma_1^2/\sigma_W^2$ is the reliability ratio associated with $\overline{W}_j$ (Carroll et al., 2006, Section 3.2.1), in which σ_1^2 and σ_W^2 denote the variance of X_1 and that of W, respectively. Replacing each unknown quantity in $E(X_{1,j}|\overline{W}_j)$ with its method-of-moments estimator yields an "estimator" or prediction of $X_{1,j}$ given by

$$\hat{X}_{1,j}^* = \overline{W} + \hat{\lambda}(\overline{W}_j - \overline{W}), \text{ for } j = 1, \ldots, n, \tag{2.4.4}$$

where $\overline{W} = n^{-1}\sum_{j=1}^n \overline{W}_j$ and $\hat{\lambda} = \hat{\sigma}_1^2/\hat{\sigma}_W^2$, in which $\hat{\sigma}_W^2$ is the sample variance of $(\overline{W}_1, \ldots, \overline{W}_n)$, $\hat{\sigma}_1^2 = (\hat{\sigma}_W^2 - \hat{\sigma}_u^2)_+$, and $\hat{\sigma}_u^2 = n^{-1}\sum_{j=1}^n S_j^2/n_j$, recalling that, for $j = 1, \ldots, n$, S_j^2 is the sample variance of $(W_{j,1}, \ldots, W_{j,n_j})$ computed earlier to evaluate the corrected score and the corrected log-likelihood. The idea of regression calibration is to regress Y on the estimated covariate $\hat{X}_1^*$ defined by (2.4.4) and $\mathbf{X}_{-1} = (X_2, \ldots, X_p)^{\mathrm{T}}$ instead of regressing on $(W, \mathbf{X}_{-1}^{\mathrm{T}})^{\mathrm{T}}$. Even though this idea often yields estimators of parameters in the primary regression model improved over naive estimators, Buonaccorsi et al. (2018) noted that (2.4.4) tends to underestimate the variability of the true covariate and thus can be problematic if used in a bootstrap procedure as we intend to. They then proposed to use

$$\hat{X}_{1,j} = \overline{W} + \hat{\lambda}^{1/2}(\overline{W}_j - \overline{W}), \text{ for } j = 1, \ldots, n, \tag{2.4.5}$$

as estimated covariate data instead so that these estimated covariate values have the mean and variance coinciding with method-of-moments estimates for the mean and variance of X_1.

With this last hurdle resolved, we are in the position to present the detailed algorithm of the parametric bootstrap for estimating the p-value associated with $\tilde{Q}(\hat{\boldsymbol{\Omega}}; \mathcal{D}^*)$ based on M bootstrap samples next.

Step 1: Fit the beta modal regression model with classical measurement error

to $\mathcal{D}^*$ by applying the MCCL method in Section 2.3.3. This gives the MCCL estimate $\hat{\boldsymbol{\Omega}} = (\hat{\boldsymbol{\beta}}^{\top}, \hat{m})^{\top}$.

Step 2: Compute the test statistic $\tilde{Q}(\hat{\boldsymbol{\Omega}}; \mathcal{D}^*)$.

For $d = 1, \ldots, M$, repeat Steps 3–5,

Step 3: For $j = 1, \ldots, n$, generate $Y_j^{(d)}$ from beta$(1+\hat{m}\hat{\theta}(\hat{X}_{1,j}, \mathbf{X}_{-1,j}),\ 1+\hat{m}\{1-\hat{\theta}(\hat{X}_{1,j}, \mathbf{X}_{-1,j})\})$, and generate $W_{j,k}^{(d)} = \hat{X}_{1,j} + U_{j,k}^{(d)}$, for $k = 1, \ldots, n_j$, where $\hat{X}_{1,j}$ is given by (2.4.5), and $\{U_{j,k}^{(d)}\}_{k=1}^{n_j}$ are i.i.d. from $N(0, S_j^2)$. Let $\widetilde{W}_j^{(d)} = \{W_{j,k}^{(d)}\}_{k=1}^{n_j}$. This yields the d-th set of bootstrap data, $\mathcal{D}^{(d)} = \{(Y_j^{(d)}, \widetilde{W}_j^{(d)}, \mathbf{X}_{-1,j})\}_{j=1}^{n}$.

Step 4: Fit the beta modal regression model with classic measurement error to $\mathcal{D}^{(d)}$, and obtain the MCCL estimate of $\boldsymbol{\Omega}$, denoted by $\hat{\boldsymbol{\Omega}}^{(d)}$.

Step 5: Compute the test statistic, $\tilde{Q}(\hat{\boldsymbol{\Omega}}^{(d)}; \mathcal{D}^{(d)})$.

Step 6: Estimate the p-value by $M^{-1}\sum_{d=1}^{M} I\left\{\tilde{Q}\left(\hat{\boldsymbol{\Omega}}^{(d)}; \mathcal{D}^{(d)}\right) > \tilde{Q}\left(\hat{\boldsymbol{\Omega}}; \mathcal{D}^*\right)\right\}$.

Empirical evidence from the simulation study presented in the next section suggest that the proposed bootstrap procedure can estimate the null distribution of $\tilde{Q}(\hat{\boldsymbol{\Omega}}; \mathcal{D}^*)$ accurately enough to preserve the right size of the test for model misspecification over a wide range of significance levels.

2.5 Simulation study

We carry out simulation study to inspect finite sample performance of the proposed estimation method and the diagnostic method. The source code to reproduce results in this section is publicly available on the journal's web page.

2.5.1 Design of simulation experiments

We generate data from each of the following four data generation processes.

(M1) Generate response data according to (2.2.2), with $m = 3$, $\theta(\mathbf{X}) = 1/\{1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2)\}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}} = (0.25, 0.25, 0.25)^{\mathrm{T}}$, $X_2 \sim$ Bernoulli(0.5), and $X_1 | X_2 \sim N(I(X_2 = 1) - I(X_2 = 0), 1)$, where $I(\cdot)$ is the indicator function. Contaminate data of X_1 according to (2.2.1) to generate $W_{j,k}$, for $j = 1, \ldots, n$ and $k = 1, 2, 3$, with $U_{j,k} \sim N(0, \sigma_u^2)$.

(M2) Same as (M1) except for that $m = 40$ and $\theta(\mathbf{X}) = 1/\{1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_1^2)\}$, with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}} = (1, 1, 1, 1)^{\mathrm{T}}$.

(M3) Same as (M1) except for that $\theta(\mathbf{X}) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}} = (1, 1, 1)^{\mathrm{T}}$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$.

(M4) Generate response data $\{Y_j\}_{j=1}^n$ according to $Y_j = (Y_j^* - Y_{(1)}^*)/(Y_{(n)}^* - Y_{(1)}^*)$, for $j = 1, \ldots, n$, where $Y_{(1)}^*$ and $Y_{(n)}^*$ are the minimum and maximum order statistics of data $\{Y_j^*\}_{j=1}^n$, respectively, $Y_j^* \mid \mathbf{X}_j \sim$ Gumbel$(\theta(\mathbf{X}_j), \gamma^{-1}\{1 - 2\theta(\mathbf{X}_j)\}/(2 + m))$, in which $\theta(\mathbf{X}_j) < 0.5$ is the mode formulated as that in (M1) with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (1, 1, 1)^{\mathrm{T}}$, $\gamma^{-1}\{1 - 2\theta(\mathbf{X}_j)\}/(2 + m)$ is the scale of the Gumbel distribution, and γ stands for the Euler–Mascheroni constant.

Despite the data generation process used to generate a particular data set, we always assume a beta modal regression model with $\theta(\mathbf{X})$ specified as that in (M1) when carrying out modal regression analysis of Y on $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$. By so doing, the design in (M1) allows us to monitor point estimation in the absence of model misspecification, and the latter three designs can be used to study operating characteristics of the proposed model diagnostic method in the presence of different sources of model misspecification. In particular, fitting the assumed model to data generated according to (M2) creates a scenario where one misspecifies the linear predictor in the regression function. When data are generated from (M3), the assumed model has a wrong link function. Finally, fitting the assumed model to data from (M4) gives rise to the most

severe model misspecification in the sense that the true distribution of Y given $\mathbf{X}$ is outside of the beta family.

2.5.2 Performance of point estimation

Besides assessing the quality of the MCCL estimator of $\boldsymbol{\Omega}$ in comparison with the naive maximum likelihood estimator, we aim at addressing the following three issues of point estimation in the simulation study: (i) the impact of having an error-free covariate along with an error-prone covariate on covariate effects estimation; (ii) the quality of the variance estimation based on (2.3.10); (iii) the robustness of the MCCL estimator to the normality assumption on U. We bring up the third issue because the corrected score method is developed under the assumption of normal measurement error. Due to our focus on covariate effects estimation in the presence of an error-prone covariate in a modal regression model for a bounded response, none of the existing modal regression methods accounting for measurement error referenced in Section 2.1 serves as a sensible competing method in the current simulation study (e.g., there is no covariate effect parameters $\boldsymbol{\beta}$ in a nonparametric modal regression model) .

Based on data generated according to (M1) with $\sigma_u^2 = 0.6, 1.2$, we obtain the MCCL estimate of $\boldsymbol{\Omega}$ using $B = 100$ and the naive maximum likelihood estimate that ignores measurement error in X_1. Table 2.1 provides the median of MCCL estimates $\hat{\boldsymbol{\Omega}}$ and the median of naive estimates across 1000 Monte Carlo replicates at each of the two sample sizes $n = 100, 200$. In contrast to the naive estimates that exhibit bias that do not diminish as the sample size increases, the MCCL estimates are much improved despite the severity of error contamination in X_1. Not surprisingly, the MCCL estimator corrects the bias of the naive estimator at the price of an inflation in variation.

The attenuation effect of measurement error on the naive covariate effect estimation for X_1 is evident in Table 2.1. In contrast, the covariate effect estimation for the error-

Table 2.1 Medians of MCCL estimates and medians of naive estimates across 1000 Monte Carlo replicates generated according to (M1). The number in parentheses following each median is the interquartile range of the 1000 realizations of an estimator.

		β_0	β_1	β_2	m
		$\sigma_u^2 = 0.6$			
$n = 100$	MCCL	0.23 (0.35)	0.24 (0.22)	0.26 (0.59)	3.24 (0.95)
	Naive	0.21 (0.31)	0.21 (0.19)	0.32 (0.50)	3.16 (0.92)
$n = 200$	MCCL	0.24 (0.23)	0.25 (0.15)	0.26 (0.40)	3.11 (0.68)
	Naive	0.21 (0.24)	0.21 (0.13)	0.33 (0.36)	3.07 (0.59)
		$\sigma_u^2 = 1.2$			
$n = 100$	MCCL	0.23 (0.35)	0.24 (0.24)	0.27 (0.65)	3.25 (0.96)
	Naive	0.19 (0.30)	0.18 (0.17)	0.37 (0.48)	3.15 (0.91)
$n = 200$	MCCL	0.25 (0.25)	0.25 (0.18)	0.25 (0.43)	3.13 (0.67)
	Naive	0.18 (0.23)	0.18 (0.11)	0.39 (0.36)	3.05 (0.60)

free covariate X_2 is noticeably overestimated by the naive method. One may wonder if the observed opposite directions in the bias of naive estimation of two covariates effects persists when the two covariates are independent. This relates to the first issue brought up above. To address this issue, we revise the data generating process in (M1) in that $X_1 \sim N(0, 1)$. Figure 2.1 includes boxplots of two sets of regression coefficients estimates, including the MCCL estimates and the naive estimates, under (M1) where X_1 and X_2 are dependent (see the left panel in Figure 2.1) and under the revised (M1) with X_1 and X_2 independent (see the right panel in Figure 2.1). Here, we set $n = 2000$ for each of 1000 Monte Carlo replicates. Interestingly, when X_2 is independent of the error-prone covariate X_1, naive estimation for the covariate effect of X_2 does not appear to be affected by measurement error. Regardless, the attenuation in the estimated covariate effect for X_1 remains.

Table 2.2 presents the average of standard deviation estimation of each parameter in $\mathbf{\Omega}$ based on (2.3.10) across 1000 Monte Carlo replicates from (M1) with $n = 200$. The Monte Carlo standard deviation of each parameter estimate in $\mathbf{\Omega}$ is used as a reference/gold standard in this table. The proximity of the standard deviation estimate

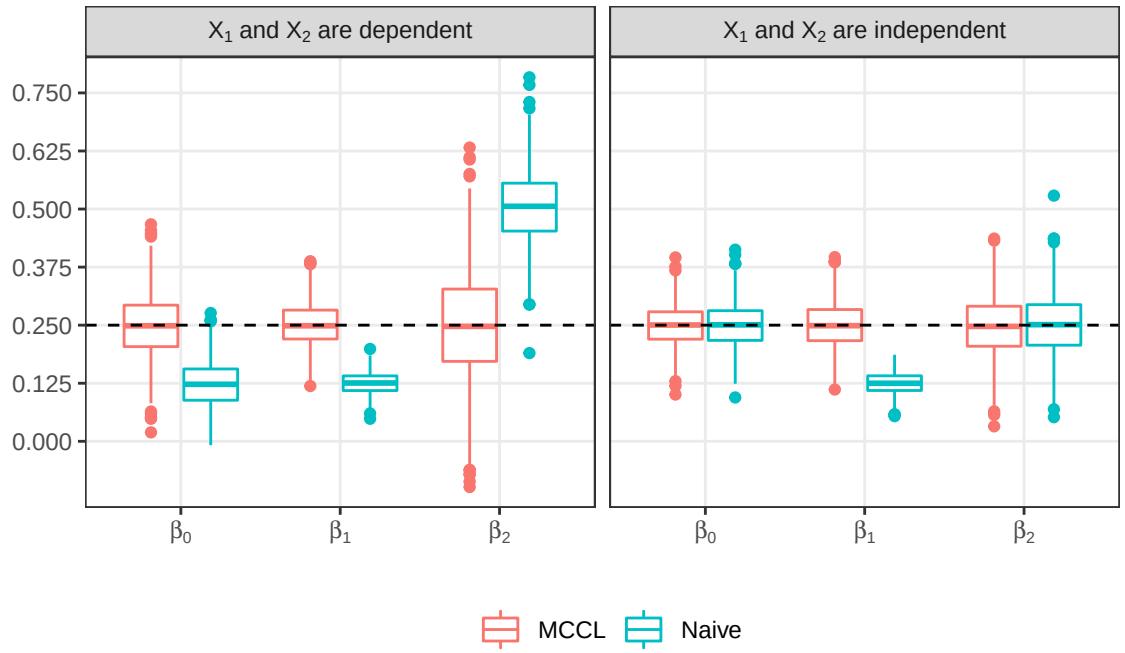


Figure 2.1 Boxplots of regression coefficients estimates under (M1) with X_1 and X_2 dependent (left panel) and those under a revised version of (M1) with X_1 and X_2 independent (right panel). The two boxes associated with each parameter correspond to two estimators (from left to right): the MCCL estimator (red box) and the naive estimator (cyan box).

with the reference shown in the table suggests that the sandwich variance estimator in (2.3.10) provides reliable estimation for the variance of the MCCL estimator. This settles the second issue.

Table 2.2 Averages of standard deviation estimates, $\widehat{\text{s.d.}}$, and empirical standard deviation, s.d., across 1000 Monte Carlo replicates from (M1) with $\sigma_u^2 = 1.2$ and $n = 200$. Numbers in parentheses are Monte Carlo standard errors associated with the Monte Carlo means.

	β_0		β_1		β_2		m	
	$\widehat{\text{s.d.}}$	s.d.	$\widehat{\text{s.d.}}$	s.d.	$\widehat{\text{s.d.}}$	s.d.	$\widehat{\text{s.d.}}$	s.d.
MCCL	0.19 (0.03)	0.19	0.13 (0.03)	0.13	0.32 (0.06)	0.32	0.48 (0.07)	0.51
Naive	0.16 (0.02)	0.17	0.09 (0.01)	0.09	0.26 (0.03)	0.26	0.47 (0.05)	0.48

The third issue concerns the normality assumption on measurement error in the development of the Monte Carlo corrected score method. To assess the robustness

of the MCCL estimator to this normality assumption, we revise (M1) by letting $U_{j,k} \sim \text{Laplace}(0, 0.5^{1/2})$ instead and set $n = 200$. Table 2.3 provides summary statistics of parameter estimates as those shown in Table 2.1 under this revised setting. As one can see from Table 2.3, despite the violation of the normality assumption on U, the MCCL estimates remain close to the truth and significantly outperform the naive estimates. This robustness feature of the Monte Carlo corrected score method is also noted and explained in Novick and Stefanski (2002).

Table 2.3 Medians of MCCL estimates and medians of naive estimates across 1000 Monte Carlo replicates generated according to (M1) with $U_{j,k} \sim \text{Laplace}(0, 0.5^{1/2})$ and $n = 200$. The number in parentheses following each median is the interquartile range of the 1000 realizations of an estimator.

	β_0	β_1	β_2	m
MCCL	0.24 (0.18)	0.24 (0.17)	0.25 (0.27)	3.08 (0.64)
Naive	0.12 (0.20)	0.13 (0.10)	0.48 (0.32)	3.05 (0.60)

2.5.3 Performance of the model diagnostic method

Using 5000 Monte Carlo replicates from (M1) with $\sigma_u^2 = 1.2$ at each sample size level in $n = 100$, 200, 500, 1000, we implement the bootstrap algorithm related in Section 2.4 with $M = 300$ bootstrap samples to obtain estimated p-values associated with the test statistic $\tilde{Q}(\hat{\boldsymbol{\Omega}}; \mathcal{D}^*)$. We then record the proportion of replicates, across 5000 replicates, that lead to rejection of the null hypothesis of no model misspecification at various nominal levels. This rejection rate can be viewed as an empirical size of the test at a pre-specified significance level. Figure 2.2 depicts this rejection rate versus the significance level, from which one can see that the size of the test is well controlled by the bootstrap procedure over a wide range of nominal levels.

Table 2.4 presents rejection rates of the model diagnostic method in the presence of different forms of model misspecification that occur when fitting data generated according to (M2)–(M4) while assuming a beta modal regression model specified in

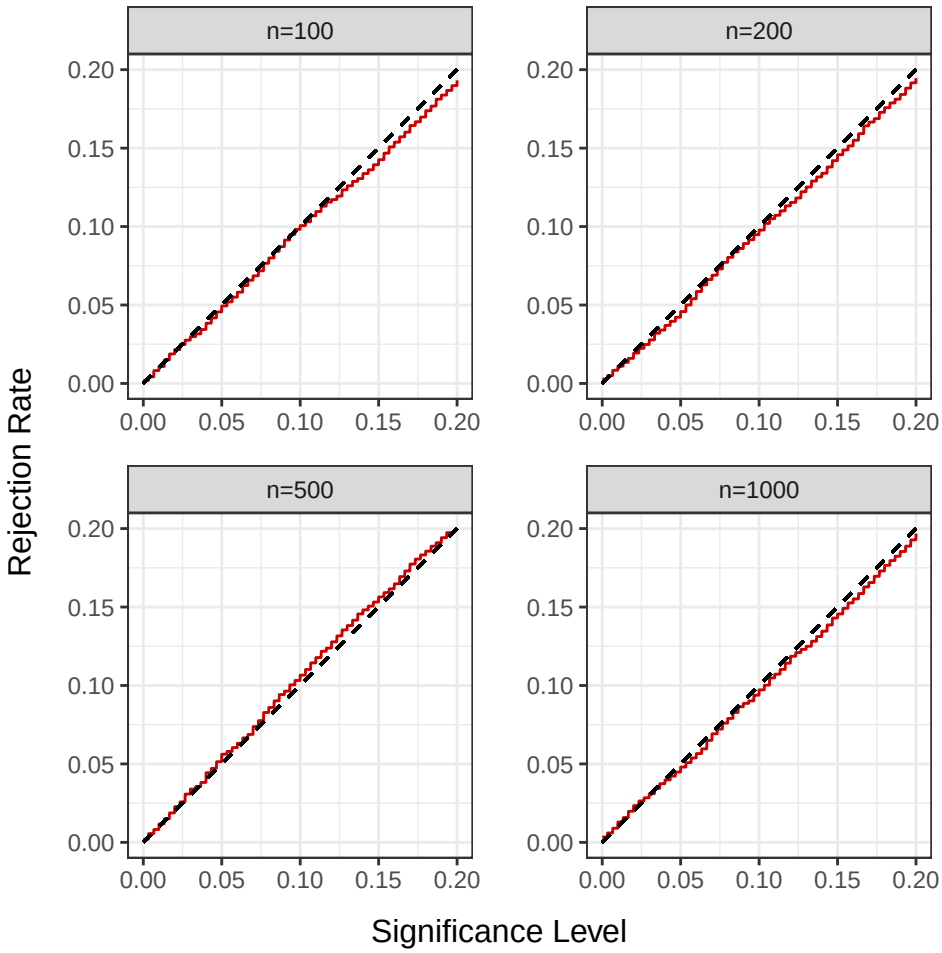


Figure 2.2 Rejection rates associated with the score-based diagnostic test across 5000 Monte Carlo replicates from (M1) versus the nominal level of the test. Black dashed lines are the 45° reference lines.

(M1). As one can see in Table 2.4, the proposed score-based test has moderate power to detect a misspecified form of the linear predictor, with the power steadily increasing as n increases, and is especially powerful in detecting violation of the distributional assumption on Y given covariates; but the test is less sensitive to link misspecification. Low power of most goodness-of-fit tests to detect link misspecification have been reported in the context of generalized linear models (e.g., Hosmer et al., 1997). Given these reported findings in the literature, the low power observed under design (M3) may not be surprising, especially with the high similarity of the logit link in the assumed model with the probit link in the true model in (M3).

Table 2.4 Rejection rates of the score-based diagnostic test resulting from 300 Monte Carlo replicates in the presence of four types of model misspecification in (M2)–(M4)

Model	$n = 200$	$n = 300$	$n = 400$	$n = 500$
(M2)	0.277	0.417	0.557	0.600
(M3)	0.100	0.107	0.143	0.123
(M4)	0.997	1.000	0.993	1.000

2.6 Real-life data application

In this section, we analyze data arising from two different applications where a covariate of interest cannot be observed directly. Besides dealing with scientific questions in relevant fields, these applications provide opportunities for us to address some practical issues one faces when implementing the proposed estimation method and diagnostic method not discussed in the simulation study.

2.6.1 Application to dietary data

Food Frequency Questionnaire (FFQ) is a convenient and inexpensive dietary assessment instrument in epidemiologic studies. To study the association between an individual's FFQ intake and his/her long-term usual intake as the univariate covariate X, we analyze a dietary data set from Women's Interview Survey of Health (Carroll et al., 1997). The data set contains 271 females' FFQ intake records, measured as the percentage calories from fat, and six 24-hour food recalls, $W_{j,k}$, for $j = 1, \ldots, 271$ and $k = 1, \cdots, 6$. Because the j-th subject's long-term usual intake X_j cannot be measured directly, a generally accepted practice in epidemiology is to use $\overline{W}_j = \sum_{k=1}^{6} W_{j,k}/6$ as a surrogate of X_j, for $j = 1, \ldots, 271$. According to the preliminary analysis in existing literature, the distribution of the FFQ intake appears to be right-skewed and potentially heavy-tailed, which motivates the consideration of a modal regression model in place of a mean regression model. Here, we assume a beta modal regression model given in (2.2.2) with $\theta(X) = 1/\{1 + \exp(-\beta_0 - \beta_1 X)\}$ for the response data

$\{Y_j\}_{j=1}^{271}$, where Y_j is the j-th subject's FFQ intake in kilocalorie divided by 8000, a biologically plausible upper bound of daily energy intakes for a general population.

We obtain the MCCL estimate of $\boldsymbol{\Omega} = (\beta_0, \beta_1, \log m)^{\mathrm{T}}$ according to (2.3.9), and also carry out regression analysis that ignores measurement error to obtain a naive maximum likelihood estimate of $\boldsymbol{\Omega}$. These two sets of estimates are given in Table 2.5. The covariate effect associated with the long-term intake suggested by the naive estimate is substantially weaker than that indicated by the MCCL estimate, implying potentially significant attenuation on the covariate effect due to measurement error in the former, whereas the latter corrects for this attenuation. Figure 2.3 depicts the estimated regression functions $\hat{\theta}(x)$ resulting from these two methods, imposed on the scaled response data versus the surrogate covariate data. This pictorial contrast between the two estimated regression functions shows that the proposed method is able to capture the underlying positive non-linear covariate effect that is partially concealed or weakened by the naive method. Finally, applying the proposed diagnostic method to this data set with $M = 300$ bootstrap samples yields an estimated p-value of 0.61. We thus conclude lack of sufficient data evidence to indicate the assumed beta modal regression model inadequate for this application.

Table 2.5 Estimates of parameters in the beta modal regression model applied to the dietary data, along with the corresponding estimated standard errors in parentheses

Method	β_0	β_1	$\log m$
MCCL	-1.569 (0.049)	1.056 (0.420)	3.298 (0.345)
Naive	-1.581 (0.041)	0.270 (0.058)	2.979 (0.094)

2.6.2 Application to Alzheimer's disease data

Medical researchers have long recognized that cerebral atrophy is associated with dementia, and extensive research have been conducted to understand the association between volumetric changes of different brain regions with the severity of dementia. Abundant data collected from this line of research are available in the Alzheimer's

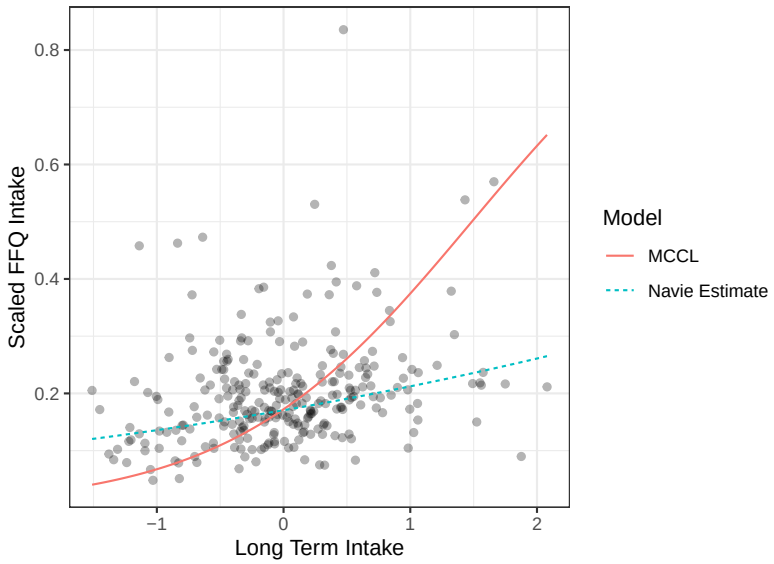


Figure 2.3 Estimated conditional mode functions for the dietary data based on the MCCL estimate (red solid line) and the naive estimate (cyan dashed line), respectively. Observed covariate data $\{\overline{W}_j\}_{j=1}^{271}$ are treated as surrogates of long-term usual intakes in the scatter plot of the observed data (solid dots).

Disease Neuroimaging Initiative (ADNI) database (`http://adni.loni.usc.edu/`). Zhou and Huang (2020) analyzed a data set relating to 245 individuals diagnosed with mild cognitive impairment from this database. The goal is to study roles that an individual's volumetric measure of entorhinal cortex (ERC) and that of hippocampus (HPC) play in predicting one's risk of developing Alzheimer's disease. An individual's test score from the Alzheimer's disease assessment scale, known as ADAS-11, at month 12 since entering the ADNI cohort is used to assess one's severity of cognitive impairment. Covariates of interest are the volumetric change in ERC (ERC.change) and that in HPC (HPC.change) at month 12 compared to the baseline measures collected at month 6. Assuming these volumetric measures are observed precisely, Zhou and Huang (2020) fitted the data to the beta modal regression model for the response Y defined as an individual's ADAS-11 score divided by a perfect score of 70, with the log-log link in the mode function, $\theta(\mathbf{X}) = \exp\{-\exp(-\beta_0 - \beta_1 \times \text{ERC.change} - \beta_2 \times \text{HPC.change})\}$, and showed that it provides a better fit for the data compared to the beta mean regression model proposed by Ferrari and Cribari-Neto

(2004).

In reality, measuring ERC volume is challenging because of lateral border discrimination from the perirhinal cortex (Price et al., 2010), and the accuracy of HPC measurements is also in question (Maclaren et al., 2014). It is thus more sensible to view the observed volumetric change of ERC or that of HPC as a noisy surrogate of the actual amount of change. Despite of which covariate is viewed as error-prone, the current data present some challenges due to the lack of replicate measures for an individual's true covariate value, and thus the estimation methods proposed in Section 2.3 are not applicable. For example, in (2.3.8), the term multiplying the imaginary unit i is equal to zero now with the number of replicates $n_j = 1$, making the "corrected" log-likelihood the same as the naive log-likelihood. A quick fix to the problem is to invoke a similar strategy of correcting naive scores to account for measurement error as discussed in Novick and Stefanski (2002). Following this strategy, a corrected log-likelihood evaluated at the j-th data point to use in place of (2.3.8) is

$$\tilde{\ell}(\boldsymbol{\Omega}; Y_j, \mathbf{W}_j, \tilde{\mathbf{Z}}_j) = \frac{1}{B} \sum_{b=1}^{B} \mathrm{Re}\{\ell(\boldsymbol{\Omega}; Y_j, \mathbf{W}_j + i\boldsymbol{\Sigma}_u^{1/2} \mathbf{Z}_{j,b})\}, \tag{2.6.1}$$

where $\tilde{\mathbf{Z}}_j = \{\mathbf{Z}_{j,b}\}_{b=1}^{B}$, for $j = 1, \ldots, n$, and $\{\mathbf{Z}_{j,b}, b = 1, \ldots, B\}_{j=1}^{n}$ are independent p-dimensional normal random vectors with mean zero and variance-covariance as an identity matrix, which accommodates multiple error-prone covariates in $\mathbf{X}$ by letting $\mathbf{W}_j$ be a p-dimensional multivariate surrogate of $\mathbf{X}_j$, contaminated by a multivariate normal measurement error $\mathbf{U}_j$ with variance-covariance matrix $\boldsymbol{\Sigma}_u$. By setting all entries in $\boldsymbol{\Sigma}_u$ at zero except for the first diagonal entry gives rise to the case considered in the majority of this chapter with only X_1 prone to error. Certainly, not having replicate measures still creates an obstacle to implementing this strategy due to its dependence on $\boldsymbol{\Sigma}_u$ that cannot be estimated without replicate measures of a true multivariate covariate value or other external validation data. A well-accepted practice among statisticians in similar situations is to carry out sensitivity analysis where one analyzes the data under different assumptions for the parameter, such as $\boldsymbol{\Sigma}_u$ in our

case, that one lacks data information to infer. If one obtains drastically different inference results when assuming different values for $\mathbf{\Sigma}_u$, including a matrix of zeros corresponding to naive estimation that ignores measurement error, then one may recommend to exercise caution when interpreting results from an inference procedure that assumes error-free covariates.

For illustration purposes, we assume in the sensitivity analysis four values for $\mathbf{\Sigma}_u$ listed in Table 2.6, where inference results for model parameters under each assumed $\mathbf{\Sigma}_u$ are provided. According to Table 2.6, all four rounds of regression analyses lead to the conclusion that the volumetric change of ERC is an influential predictor for the severity of cognitive impairment, even though the magnitude of the estimated covariate effect is sensitive to the assumed error variance associated this covariate. In particular, when assuming imprecise measurements for ERC.change, the revised MCCL method that employs the corrected log-likelihood in (2.6.1) with $B = 1000$ produces results indicating a much stronger association than the naive analysis. By comparison, the magnitude of the estimate for the HPC.change effect is less sensitive to the assumed $\mathbf{\Sigma}_u$, but its statistical significance is noticeably affected by it. For example, one would conclude a moderately significant covariate effect of HPC.change based on the naive analysis assuming error-free covariates, but claim a highly significant, or moderately significant, or nonsignificant HPC.change effect depending on which covariate(s) one assumes to be error-prone and the severity of error contamination. This phenomenon is a reminiscence of an observation made in Figure 2.1, and may suggest that ERC.change and HPC.change are correlated. In fact, measurements of ERC and HPC via magnetic resonance imaging are known to be highly correlated with observed clinical alterations in patients suffering mild cognitive impairment or at dementia phases of Alzheimer's disease (Desikan et al., 2010; Jack et al., 2013; Varon et al., 2014).

In conclusion, results from the sensitivity analysis suggest that volumetric measures

Table 2.6 Sensitivity analysis using the ADNI data for the beta modal regression with the log-log link. Numbers in parentheses are estimated standard errors. Numbers in square brackets are p-values associated with covariate effects.

$\boldsymbol{\Sigma}_u$	β_0	β_1 (ERC.change)	β_2 (HPC.change)	$\log m$
$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	-0.69 (0.03)	-0.12 (0.05) [0.007]	-0.22 (0.11) [0.054]	2.78 (0.15)
$\begin{bmatrix} 0.16 & 0 \\ 0 & 0 \end{bmatrix}$	-0.83 (0.09)	-2.47 (0.00) [0.000]	-0.30 (1.03) [0.773]	2.40 (0.02)
$\begin{bmatrix} 0 & 0 \\ 0 & 0.0225 \end{bmatrix}$	-0.71 (0.03)	-0.11 (0.05) [0.013]	-0.46 (0.27) [0.084]	2.80 (0.16)
$\begin{bmatrix} 0.16 & 0 \\ 0 & 0.0225 \end{bmatrix}$	-0.81 (0.03)	-2.45 (0.00) [0.000]	-0.87 (0.06) [0.000]	3.94 (0.01)

of different brain regions are likely to be subject to measurement error, and statistical analyses under the assumption of precisely measured covariates should be interpreted with caution. If replicate data are available for covariates of interest, the MCCL method can provide more reliable inference. Lastly, even though one can mimic (2.6.1) to construct a corrected score in place of $\tilde{\mathbf{S}}(\boldsymbol{\Omega}; Y_j, \widetilde{W}_j, \widetilde{T}_j, \mathbf{X}_{-1,j})$ in (2.4.3) and then formulate the test statistic $\tilde{Q}(\hat{\boldsymbol{\Omega}}; \mathcal{D}^*)$ for model diagnostics, the dependence of the revised score on the unknown $\boldsymbol{\Sigma}_u$ remains an obstacle that hinders one from using the bootstrap procedure outlined in Section 2.4 to assess statistical significance of the revised test statistic. Alternative diagnostic methods that do not rely on parametric bootstrap or corrected score (e.g. Huang et al., 2006) can be used to detect inadequate assumptions imposed on the primary regression model.

2.7 Discussion

We propose an inference procedure based on the idea of corrected score that falls in the framework of M-estimation for modal regression with an error-prone covariate. Even though in this chapter we focus on the beta modal regression model as the primary regression model, the proposed MCCL method is applicable in other parametric modal regression models, such as the gamma modal regression models for non-negative

responses proposed by Aristodemou (2014) and Bourguignon et al. (2020). A Python package for implementing the proposed methods for beta modal regression with errors-in-covariate is available at `https://pypi.org/project/pybetareg/`. All computer programs used in this chapter is available at `https://github.com/rh8liuqy/Modal_regression_with_measurement_error`.

To accommodate situations without replicate measures of the true covariate or settings with multiple error-prone covariates, the MCCL method can be easily revised as demonstrated in Section 2.6.2, although one needs to specify the variance (or the variance-covariance matrix) of the (vector-valued) measurement error if one lacks replicate data or external validation data to estimate it.

Focusing on the current beta modal regression models, some extensions are worthy of further investigation, such as a zero-inflated beta modal regression model to fit disease prevalence data especially suitable for rare diseases. Another follow-up research direction is variable selection based on a parametric modal regression model with or without measurement error contamination in covariates.

Chapter 3

The Flexible Gumbel Distribution

3.1 Introduction

When data contain outliers that cause heavy tails or are potentially skewed, the mode is a more sensible representation of the central location of data than the mean or median. The timely review on mode estimation and its application by Chacón (2020) and references therein provide many examples in various fields of research where the mode serves as a more informative representative value of data. Most existing methods developed to draw inference for the mode are semi-/non-parametric in nature, starting from early works on direct estimation in the 1960s (Chernoff, 1964; Dalenius, 1965; Venter, 1967) to more recent works based on kernel density estimation (Chen, 2018) and quantile-based methods (Ota et al., 2019; Zhang et al., 2021). There are two main reasons contributing to the long-lasting trend of opting to semi-/non-parametric methods for mode estimation, despite the fact that inference procedures proposed along these veins are usually less straightforward to implement (e.g., involving bandwidth selection), and less efficient than their parametric counterparts. First, a parametric model typically imposes stringent constraints on the relationship between the mode and other location parameters that may not be satisfied in a given application. Second, very few existing named distribution families that allow inclusion of both symmetric and asymmetric distributions in the same family can be parameterized so that it is indexed by the mode as the location parameter along with other parameters, such as shape or scale parameters. In this study, we alleviate concerns raised by both reasons

that discourage use of parametric methods for mode estimation by formulating a flexible distribution indexed by the (unique) mode and parameters controlling the shape and scale.

When it comes to modeling heavy-tailed data, the Gumbel distribution (Gumbel, 1941) is arguably one of the most widely used models in many disciplines. Indeed, as a case of the generalized extreme value distribution (Jenkinson, 1955), the Gumbel distribution for the maximum (or minimum) is well-suited for modeling extremely large (or small) events that produce heavy-tailed data. For example, it is often used in hydrology to predict extreme rainfall and flood frequency (Loaiciga and Leipnik, 1999; Koutsoyiannis, 2004; Dawley et al., 2019). In econometrics, the Gumbel distribution plays an important role in modeling extreme movements of stock prices and large changes in interest rates (Bali, 2003; Pratiwi et al., 2019). The Gumbel distribution is indexed by the mode and a scale parameter, and thus is convenient for mode estimation. However, the Gumbel distribution for the maximum (or minimum) is right-skewed (or left-skewed) with the skewness fixed at around 1.44 (or -1.44), and the kurtosis fixed at 5.4 across the entire distribution family. Thus it may be too rigid for scenarios where the direction and extremeness of outliers presented in data are initially unclear, or when the direction and level of skewness are unknown beforehand. Constructions of more flexible distributions that overcome these limitations have been proposed. In particular, Cooray (2010) applied a logarithmic transformation on a random variable following the odd Weibull distribution to obtain the so-called generalized Gumbel distribution that includes the Gumbel distribution as a subfamily. But the mode of the generalized Gumbel distribution is not a location parameter this distribution is indexed by, or an explicit function of other model parameters. Shin et al. (2015b) considered mixture distributions with one of the components being the Gumbel distribution and the other component(s) being Gumbel of the same skewness direction or a different distribution, such as the gamma distribution. Besides the

same drawback pointed out for the generalized Gumbel distribution, it is difficult to formulate a unimodal distribution following their construction of mixtures, and thus their proposed models are unsuitable when unimodality is a feature required to make inferring the mode meaningful, such as in a regression setting, as in modal regression (Yao et al., 2012; Yao and Li, 2013; Chen, 2018).

With heavy-tailed data in mind and the mode as the location parameter of interest, we construct a new unimodal distribution that does not impose stringent constraints on how the mode relates to other central tendency measures, while allowing a range of kurtosis wide enough to capture heavy tails at either direction, as well as different degrees and directions of skewness. This new distribution, called the flexible Gumbel (FG) distribution, is presented in Section 3.2, where we study properties of the distribution and discuss identifiability of the model. We present a frequentist method and a Bayesian method for estimating parameters in the FG distribution in Section 3.3. Finite sample performance of these methods are inspected in simulation study in Section 3.4, followed by an application of the FG distribution in hydrology in Section 3.5. Section 3.6 demonstrates fitting a modal regression model based on the FG distribution to data from a criminology study. Section 3.7 highlights contributions of the study and outlines future research directions.

3.2 The flexible Gumbel distribution

The probability density function (pdf) of the Gumbel distribution for the maximum is given by

$$f(x;\theta,\sigma) = \frac{1}{\sigma}\exp\left\{-\frac{x-\theta}{\sigma} - \exp\left(-\frac{x-\theta}{\sigma}\right)\right\}, \tag{3.2.1}$$

where θ is the mode and $\sigma > 0$ is a scale parameter. The pdf of the Gumbel distribution for the minimum with mode θ and a scale parameter σ is given by

$$f(x;\theta,\sigma) = \frac{1}{\sigma}\exp\left\{\frac{x-\theta}{\sigma} - \exp\left(\frac{x-\theta}{\sigma}\right)\right\}. \tag{3.2.2}$$

We define a unimodal distribution for a random variable Y via a mixture of the two Gumbel distributions specified by (3.2.1) and (3.2.2) that share the same mode θ while allowing different scale parameters, σ_1 and σ_2, in the two components. We call the resultant distribution the flexible Gumbel distribution, FG for short, with the pdf given by

$$\begin{aligned} f(y) = w \times \frac{1}{\sigma_1} \exp\left\{-\frac{x-\theta}{\sigma_1} - \exp\left(-\frac{x-\theta}{\sigma_1}\right)\right\} + \\ (1-w) \times \frac{1}{\sigma_2} \exp\left\{\frac{x-\theta}{\sigma_2} - \exp\left(\frac{x-\theta}{\sigma_2}\right)\right\}, \end{aligned} \tag{3.2.3}$$

where $w \in [0,1]$ is the mixing proportion parameter. Henceforth, we state that $Y \sim \text{FG}(\theta, \sigma_1, \sigma_2, w)$ if Y follows the distribution specified by the pdf in (3.2.3).

For each component distribution of FG, the mean and median are both some simple shift of the mode, with each shift solely determined by the scale parameter. Because the two components in (3.2.3) share a common mode θ, the mode of Y is also θ, and thus the FG distribution is convenient to use when one aims to infer the mode as a central tendency measure, or to formulate parametric modal regression models (Bourguignon et al., 2020; Zhou and Huang, 2020, 2022). One can easily show that the mean of Y is $E(Y) = w(\theta + \sigma_1\gamma) + (1-w)(\theta - \sigma_2\gamma) = \theta + \{w(\sigma_1 + \sigma_2) - \sigma_2\}\gamma$, where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant. Thus the discrepancy between the mode and the mean of FG depends on three other parameters that control the scale and shape of the distribution. The median of Y, denoted by m, is the solution to the following equation,

$$w \exp\left\{-\exp\left(-\frac{m-\theta}{\sigma_1}\right)\right\} + (1-w)\left[1 - \exp\left\{-\exp\left(\frac{m-\theta}{\sigma_2}\right)\right\}\right] = 0.5.$$

Even though this equation cannot be solved for m explicitly to reveal the median in closed form, it is clear that $m - \theta$ also depends on all three other parameters of FG. In conclusion, the relationships between the three central tendency measures of FG are more versatile than those under a Gumbel distribution for the maximum or a Gumbel distribution for the minimum.

The variance of Y is $V(Y) = \{w\sigma_1^2 + (1-w)\sigma_2^2\}\pi^2/6 + w(1-w)(\sigma_1+\sigma_2)^2\gamma^2$, which does not depend on the mode parameter θ. Obviously, by setting $w = 0$ or 1, $\text{FG}(\theta, \sigma_1, \sigma_2, w)$ reduces to one of the Gumbel components. Unlike a Gumbel distribution that only has one direction of skewness at a fixed level (of ± 1.44), an FG distribution can be left-skewed, or right-skewed, or symmetric. More specifically, with the mode fixed at zero when studying the skewness and kurtosis of FG, one can show that the third central moment of Y is given by

$$w\bar{w}(\sigma_1+\sigma_2)^2\gamma\left\{\gamma^2(\bar{w}-w)(\sigma_1+\sigma_2) + 0.5\pi^2(\sigma_1-\sigma_2)\right\} + 2\zeta(3)\left(w\sigma_1^3 - \bar{w}\sigma_2^3\right), \quad (3.2.4)$$

where $\bar{w} = 1-w$, and $\zeta(3) \approx 1.202$ is Apéry's constant. Although the direction of skewness is not immediately clear from (3.2.4), one may consider a special case with $w = 0.5$ where (3.2.4) reduces to $(\sigma_1-\sigma_2)\{\gamma\pi^2(\sigma_1+\sigma_2)^2/8 + \zeta(3)(\sigma_1^2+\sigma_1\sigma_2+\sigma_2^2)\}$. Now one can see that $\text{FG}(\theta, \sigma_1, \sigma_2, 0.5)$ is symmetric if and only if $\sigma_1 = \sigma_2$, and it is left-skewed (or right-skewed) when σ_1 is less (or greater) than σ_2. The kurtosis of Y can also be derived straightforwardly, with a more lengthy expression than (3.2.4) that we omit here, which may not shed much light on its magnitude except for that it varies as the scale parameters and the mixing proportion vary, instead of fixing at 5.4 as for a Gumbel distribution. An R Shiny app depicting the pdf of $\text{FG}(\theta, \sigma_1, \sigma_2, w)$ with user-specified parameter values is available at `https://qingyang.shinyapps.io/gumbel_mixture/`, created and maintained by the first author. Along with the density function curve, the Shiny app provides skewness and kurtosis of the depicted FG density. From there one can see that the skewness can be much lower than -1.44 or higher than 1.44, and the kurtosis can be much higher than 5.4, suggesting that inference based on FG can be more robust to outliers than when a Gumbel distribution is assumed for data at hand, without imposing stringent assumption on the skewness of the underlying distribution.

The flexibility of a mixture distribution usually comes with concerns relating to identifiability (Teicher, 1961, 1963; Yakowitz and Spragins, 1968). In particular,

there is the notorious issue of label switching when fitting a finite mixture model (Redner and Walker, 1984). Take the family of two-component normal mixture (NM) distributions as an example, defined by $\{\text{NM}(\mu_1, \sigma_1, \mu_2, \sigma_2, w) : w\mathcal{N}(\mu_1, \sigma_1^2) + (1-w)\mathcal{N}(\mu_1, \sigma_2^2),\ \text{for } \sigma_1, \sigma_2 > 0 \text{ and } w \in [0,1]\}$. When fitting a data set assuming a normal mixture distribution, one cannot distinguish between, for instance, $\text{NM}(1,2,3,4,0.2)$ and $\text{NM}(3,4,1,2,0.8)$, since the likelihood of the data is identical under these two mixture distributions. As another example, for data from a normal distribution, a two-component normal mixture with two identical normal components and an arbitrary mixing proportion $w \in [0,1]$ leads to the same likelihood, and thus w cannot be identified. Teicher (1963) showed that imposing an lexicographical order for the normal components resolves the issue of non-identifiability, which also excludes mixtures with two identical components in the above normal mixture family. Unlike normal mixtures of which all components are in the same family of normal distributions, the FG distribution results from mixing two components from different families, i.e., a Gumbel distribution for the maximum and a Gumbel distribution for the minimum, with weight w on the former component. By construction, FG does not have the label-switching issue. And, according to Teicher (1963, Theorem 1), the so-constructed mixture distribution is always identifiable even when the true distribution is a (one-component) Gumbel distribution.

3.3 Statistical inference

3.3.1 Frequentist inference method

Based on a random sample of size n from the FG distirbution, $\mathbf{y} = \{y_i\}_{i=1}^n$, maximum likelihood estimators (MLE) of all model parameters in $\boldsymbol{\Omega} = (\theta, \sigma_1, \sigma_2, w)$ can be obtained via the expectation-maximization (EM) algorithm (Dempster et al., 1977). To apply the EM algorithm, we introduce a latent variable Z that follows Bernoulli(w)

such that the joint likelihood of (Y, Z) is

$$f_{Y,Z}(y, z) = \{w f_1(y; \theta, \sigma_1)\}^z \{(1-w) f_2(y; \theta, \sigma_2)\}^{1-z},$$

where $f_1(y; \theta, \sigma_1)$ is the pdf in (3.2.1) evaluated at y with the scale parameter $\sigma = \sigma_1$, and $f_2(y; \theta, \sigma_2)$ is the pdf in (3.2.2) evaluated at y with the scale parameter $\sigma = \sigma_2$. A random sample of size n from Bernoulli(w), $\mathbf{z} = \{z_i\}_{i=1}^n$, is viewed as missing data, and $\{(y_i, z_i)\}_{i=1}^n$ are viewed as the complete data in the EM algorithm. The complete-data log-likelihood is then

$$\ell(\mathbf{\Omega}; \mathbf{y}, \mathbf{z}) = \sum_{i=1}^{n} \{z_i \log(w f_1(y_i; \theta, \sigma_1)) + (1 - z_i) \log((1-w) f_2(y_i; \theta, \sigma_2))\}. \quad (3.3.1)$$

Starting from an initial estimate of $\mathbf{\Omega}$ (at the zero-th iteration), denoted by $\mathbf{\Omega}^{(0)}$, one iterates two steps referred to as the E-step and the M-step until a convergence criterion is met. In the E-step at the $(t+1)$-th iteration, one computes the conditional expectation of (3.3.1) given $\mathbf{y}$ while assuming the true parameter value to be $\mathbf{\Omega}^{(t)} = (\theta^{(t)}, \sigma_1^{(t)}, \sigma_2^{(t)}, w^{(t)})$, that is, $\mathrm{E}_{\mathbf{\Omega}^{(t)}}\{\ell(\mathbf{\Omega}; \mathbf{y}, \mathbf{z})|\mathbf{y}\}$. This conditional expectation can be shown to be

$$Q\left(\mathbf{\Omega} \left| \mathbf{\Omega}^{(t)} \right.\right) = \sum_{i=1}^{n} \left\{ T_i^{(t)} \log(w f_1(y_i; \theta, \sigma_1)) + \left(1 - T_i^{(t)}\right) \log((1-w) f_2(y_i; \theta, \sigma_2)) \right\}, \quad (3.3.2)$$

where

$$T_i^{(t)} = \mathrm{E}_{\mathbf{\Omega}^{(t)}}(Z|Y = y_i) = \frac{w^{(t)} f_1(y_i; \theta^{(t)}, \sigma_1^{(t)})}{w^{(t)} f_1(y_i; \theta^{(t)}, \sigma_1^{(t)}) + (1 - w^{(t)}) f_2(y_i; \theta^{(t)}, \sigma_2^{(t)})}. \quad (3.3.3)$$

In the M-step at the $(t+1)$-th iteration, one maximizes $Q(\mathbf{\Omega}|\mathbf{\Omega}^{(t)})$ with respect to $\mathbf{\Omega}$ to obtain an updated estimate $\mathbf{\Omega}^{(t+1)}$.

Our experience with the above EM algorithm for fitting the FG distribution suggests that maximizing $Q(\mathbf{\Omega}|\mathbf{\Omega}^{(t)})$ in (3.3.2) can be numerically challenging. We thus exploit the expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993), which replaces the M-step with a sequence of simpler conditional maximizations

referred to as the CM-step. More specifically, in the CM-step, the updating formula for w is simply $w^{(t+1)} = \sum_{i=1}^{n} T_i^{(t)}/n$. There is no closed-form updating formula for the other three parameters in $\boldsymbol{\Omega}$, but they can now be easily updated by most well-accepted one-dimensional optimization algorithms, such as the Newton-Raphson method. To ensure convergence at the global maximum, as recommended by Wu (1983), one should implement the ECM algorithm several rounds with different starting values $\boldsymbol{\Omega}^{(0)}$.

After obtaining the MLE of $\boldsymbol{\Omega}$, denoted by $\hat{\boldsymbol{\Omega}}$, we propose to use the sandwich variance estimator (Boos and Stefanski, 2013, Chapter 7) to estimate the variance-covariance matrix of $\hat{\boldsymbol{\Omega}}$. One may also estimate the variance-covariance of $\hat{\boldsymbol{\Omega}}$ based on the observed information matrix as described in Louis (1982) and Oakes (1999). The benefit of using the sandwich variance estimator is its robustness to model misspecification. Finally, the EM and ECM algorithms bear a strong resemblance to data augmentation (Wei and Tanner, 1990) in the Bayesian framework, which we turn to next for inferring $\boldsymbol{\Omega}$.

3.3.2 Bayesian inference method

In the Bayesian framework, we formulate hierarchical models starting with the FG distribution,

$$Y|\theta, \sigma_1, \sigma_2, w \sim \text{FG}(\theta, \sigma_1, \sigma_2, w),$$

followed by independent weakly informative or non-informative priors for elements in $\boldsymbol{\Omega}$,

$$\begin{aligned} \theta &\sim \mathcal{N}(0, 10^4), \\ \sigma_j &\sim \text{inv-Gamma}(1, 1), \text{ for } j = 1, 2, \\ w &\sim \text{Beta}(1, 1), \end{aligned}$$

where inv-Gamma refers to the inverse Gamma distribution. We choose the above prior for the scale parameters by following the prior selection for variance parameters suggested in Gelman (2006)

We employ the Metropolis-within-Gibbs sampler (Müller, 1991, 1993) to obtain an estimate of $\mathbf{\Omega}$ from the posterior distribution of $\mathbf{\Omega}$ given observed data $\mathbf{y}$. Similar to the EM/ECM algorithm in Section 3.3.1, the latent variable Z is also introduced as a device to carry out data augmentation. And the iterative algorithm presented next is based on the following two conditional distributions that can be easily proved,

$$z_i|\theta,\sigma_1,\sigma_2,w,\mathbf{z}_{-i},\mathbf{y} \sim \text{Bernoulli}\left(\frac{wf_1(y_i;\theta,\sigma_1)}{wf_1(y_i;\theta,\sigma_1)+(1-w)f_2(y_i;\theta,\sigma_2)}\right),$$
$$w|\theta,\sigma_1,\sigma_2,\mathbf{z},\mathbf{y} \sim \text{Beta}\left(1+\sum_{i=1}^{n} z_i,\ n+1-\sum_{i=1}^{n} z_i\right),$$

where $\mathbf{z}_{-i}$ results from dropping z_i from $\mathbf{z}$, and the first result above is also from which (3.3.3) is deduced.

The Metropolis-within-Gibbs sampler at the $(t+1)$-th iteration involves four steps outlined below.

- Step 1: For $i = 1,\ldots,n$, draw $z_i^{(t+1)}$ from Bernoulli$(T_i^{(t)})$, where $T_i^{(t)}$ is given in (3.3.3).
- Step 2: Draw $w^{(t+1)}$ from Beta $\left(1+\sum_{i=1}^{n} z_i^{(t+1)},\ n+1-\sum_{i=1}^{n} z_i^{(t+1)}\right)$.
- Step 3: Draw $\tilde{\theta}$ from $\mathcal{N}(\theta^{(t)},\tau_0)$, and update $\theta^{(t)}$ to $\theta^{(t+1)}$ according to the following decision rule,

$$\theta^{(t+1)} = \begin{cases} \tilde{\theta}, & \text{with probability } q = \min\left\{\dfrac{p(\tilde{\theta}|w^{(t+1)},\sigma_1^{(t)},\sigma_2^{(t)},\mathbf{y})}{p(\theta^{(t)}|w^{(t+1)},\sigma_1^{(t)},\sigma_2^{(t)},\mathbf{y})},\ 1\right\}, \\ \theta^{(t)}, & \text{with probability } 1-q. \end{cases}$$

- Step 4: For $j = 1,2$, draw $\tilde{\sigma}_j$ from $\mathcal{N}(\sigma_j^{(t)},\tau_j)$, and update $\sigma_j^{(t)}$ to $\sigma_j^{(t+1)}$ according

to the following decision rule, for $k \neq j$,

$$\sigma_j^{(t+1)} = \begin{cases} \tilde{\sigma}_j, & \text{with probability } q = \min\left\{ \dfrac{p(\tilde{\sigma}_j|\theta^{(t+1)}, \sigma_k^{(t)}, w^{(t+1)}, \mathbf{y})}{p(\sigma_j^{(t)}|\theta^{(t+1)}, \sigma_k^{(t)}, w^{(t+1)}, \mathbf{y})}, 1 \right\}, \\ \sigma_j^{(t)}, & \text{with probability } 1 - q. \end{cases}$$

In Steps 3 and 4, $p(\cdot|\cdot)$ refers to a conditional pdf generically, τ_0, τ_1, and τ_2 are three positive tuning parameters whose values should be chosen so that the acceptance rate at each step is around 23% (Gelman et al., 1997). To draw samples from the joint posterior distribution, there are numerous ways to design the Markov chain Monte Carlo (MCMC) sampler. Instead of the Metropolis-within-Gibbs sampler we adopt here, one may use other existing MCMC software, such as Stan (Stan Development Team, 2021), JAGS (Plummer et al., 2003), and BUGS (Spiegelhalter et al., 1996; Lunn et al., 2009), two of which we demonstrate in the Appendix. After obtaining enough high quality samples from the joint posterior distribution $p(\theta, \sigma_1, \sigma_2, w|\mathbf{y})$, Bayesian inference is straightforward, including point estimation, interval estimation, and uncertainty assessment.

3.4 Simulation study

Large-sample properties of MLEs and likelihood-based Bayesian inference under a correct model for data have been well studied. To assess finite-sample performance of the frequentist method and Bayesian method proposed in Section 3.3, we carried out simulation study with two specific aims: first, to compare inference results from the two methods; second, to compare goodness of fit for data from distributions outside of the FG family when one assumes an FG distribution and when one assumes a two-component normal mixture distribution for the data.

In the first experiment, referred to as (E1) in the sequel, we drew a random sample of size $n \in \{100, 200\}$ from an FG distribution with $\theta = 0$, $\sigma_1 = 1$, $\sigma_2 = 5$, and $w = 0.5$. Based on each simulated data set, we estimated $\boldsymbol{\Omega}$ by applying the ECM algorithm

and the Metropolis-within-Gibbs algorithm. The former algorithm produced the MLE of $\boldsymbol{\Omega}$, and we used the median of the posterior distribution of $\boldsymbol{\Omega}$ at convergence of the latter algorithm as another point estimate of $\boldsymbol{\Omega}$. Table 3.1 presents summary statistics of these estimates of $\boldsymbol{\Omega}$ and estimates of the corresponding standard deviation across 1000 Monte Carlo replicates.

Table 3.1 Frequentist and Bayesian inference results in experiment (E1) across 1000 Monte Carlo replicates. Here, point.est stands for the average of 1000 point estimates for each parameter from each method, $\widehat{\text{s.d.}}$ stands for the average of the corresponding 1000 estimated standard deviations, and s.d. refers to the empirical standard deviation of the 1000 point estimates from each method. Numbers in parentheses are 100× Monte Carlo standard errors associated with the averages.

		Frequentist			Bayesian		
sample size	parameter	point.est	$\widehat{\text{s.d.}}$	s.d.	point.est	$\widehat{\text{s.d.}}$	s.d.
$n = 100$	θ	0.002	0.198 (0.20)	0.201	0.013	0.205 (0.15)	0.203
	σ_1	0.979	0.204 (0.41)	0.216	1.014	0.224 (0.27)	0.214
	σ_2	4.932	0.590 (0.56)	0.613	4.813	0.666 (0.44)	0.615
	w	0.495	0.091 (0.09)	0.090	0.484	0.090 (0.04)	0.088
$n = 200$	θ	0.008	0.136 (0.08)	0.129	0.011	0.137 (0.07)	0.130
	σ_1	0.999	0.143 (0.21)	0.144	1.013	0.144 (0.10)	0.141
	σ_2	4.993	0.435 (0.32)	0.431	4.940	0.457 (0.20)	0.434
	w	0.500	0.064 (0.04)	0.063	0.495	0.063 (0.02)	0.062

According to Table 3.1, all estimates for parameters in $\boldsymbol{\Omega}$ are reasonably close to the truth. A closer inspection on the reported empirical mean of these estimates along with their empirical standard error suggests that, when $n = 100$, the Bayesian method may slightly underestimate σ_2, the larger of the two scale parameters of FG. We believe that this is due to the inverse gamma prior imposed on the scale parameters that is sharply peaked near zero, and thus the posterior median of the larger scale parameter tends to be pulled downwards when the sample size is not large. As the sample size increases to 200, this trend of underestimation appears to diminish. The empirical means of the standard deviation estimates from both methods are close to

the corresponding empirical standard deviations, which indicate that the variability of a point estimator is accurately estimated, whether it is based on the sandwich variance estimator in the frequentist framework, or based on the posterior sampling in the Bayesian framework. In summary, the methods proposed in Section 3.3 under both frameworks provide reliable inference for $\boldsymbol{\Omega}$ along with accurate uncertainty assessment of the point estimators when data arise from an FG distribution.

Among all existing mixture distributions, normal mixtures probably have the longest history and are most referenced in the literature. In another experiment, we compared the model fitting of normal mixture with that of FG when data arise from three heavy-tailed distributions: (E2) Laplace with the location parameter equal to zero and the scale parameter equal to 2; (E3) a mixture of two Gumbel distributions for the maximum, with a common mode at zero, scale parameters in the two components equal to 2 and 6, respectively, and the mixing proportion equal to 0.5; (E4) a t distribution with degrees of freedom equal to 5. From each of the three distributions in (E2)–(E4), we generated a random sample of size $n = 200$, following which we fit a two-component normal mixture model via the EM algorithm implemented using the R package `mixtools`, and also fit an FG model via the two algorithms described in Section 3.3. This model fitting exercise was repeated for 1000 Monte Carlo replicates under each of (E2)–(E4).

We used an empirical version of the Kullback-Leibler divergence as the metric to assess the quality of modeling fitting. We denote the true density function as $p(\cdot)$, and let $\hat{p}(\cdot)$ be a generic estimated density resulting from one of the three considered model fitting strategies. Under each setting in (E2)–(E4), a random sample of size 50000, $(x_1, \ldots, x_{50000})$, were generated from the true distribution, and an empirical version of the Kullback-Leibler divergence from $\hat{p}(\cdot)$ to $p(\cdot)$ is given by $D_{\mathrm{KL}} = (1/50000) \sum_{i=1}^{50000} \log(p(x_i)/\hat{p}(x_i))$. Figure 3.1 shows the boxplots of D_{KL} across 1000 Monte Carlo replicates corresponding to each model fitting scheme under

(E2)–(E4).

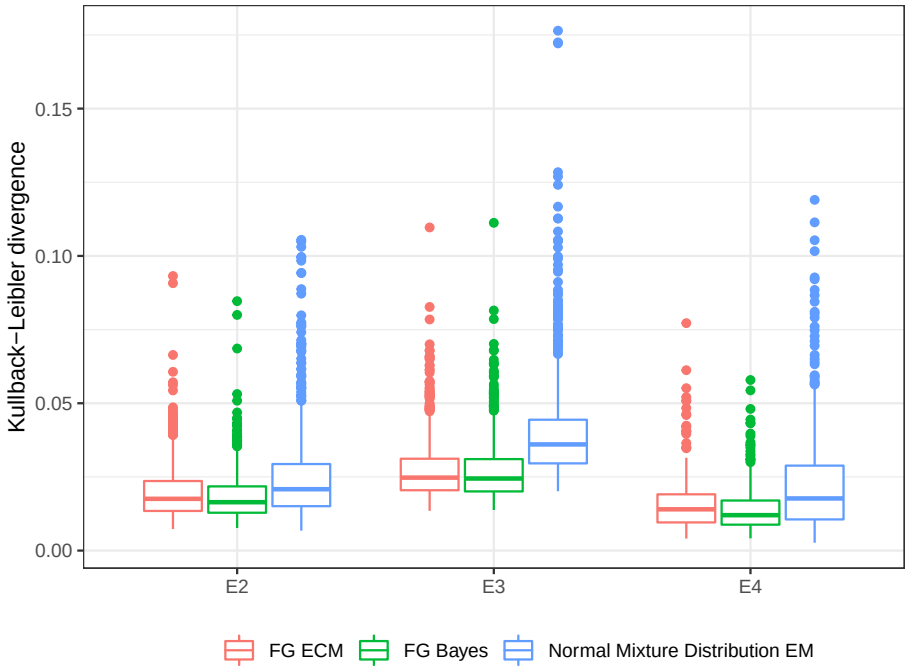


Figure 3.1 Boxplots of the empirical Kullback-Leibler divergence from an estimated density to the true density under each of the true-model settings in (E2)–(E4). Under each setting, the three considered model fitting strategies are, from left to right in the figure, (i) using the ECM algorithm to fit an FG distribution (FG ECM), (ii) using the Bayesian method to fit an FG distribution (FG Bayes), and (iii) using the EM algorithm to fit a normal mixture distribution (Normal Mixture Distribution EM).

Judging from Figure 3.1, the FG distribution clearly outperform the normal mixture when fitting data from any of the three heavy-tailed distributions in (E2)–(E4), and results from the frequentist method are comparable with those from the Bayesian method for fitting an FG model. When implementing the ECM algorithm for fitting the FG model and the EM algorithm for fitting the normal mixture, we set a maximum number of iterations at 1000. Our ECM algorithm always converged in the simulation, i.e., converged to a stationary point within 1000 iterations. But the EM algorithm for fitting a normal mixture often had trouble achieving that, with more difficulty when data come from a heavier-tailed distribution. More specifically, under (E4), which has

the highest kurtosis (equal to 9) among the three settings, the EM algorithm failed to converge in 59.9% of all Monte Carlo replicates; under (E2), which has the second highest kurtosis (equal to 6), it failed to converge in 6.7% of the replicates. Results associated with the normal mixture from these failing replicates were not included when producing the boxplots in Figure 3.1. In conclusion, the FG distribution is more suitable for symmetric or asymmetric heavy-tailed data than the normal mixture distribution.

3.5 An application in hydrology

Daily maximum water elevation changes of a waterbody, such as ocean, lake, and wetland, are of interest in hydrologic research. These changes may be close to zero in most days, but can be extremely large or small under extreme weather. From National Water Information System (`https://waterdata.usgs.gov/`), we downloaded water elevation data for Lake Murray near Columbia, South Carolina, United States, recorded from September 18, 2020 to September 18, 2021. The water elevation change of a given day was calculated by contrasting the maximum elevation and the minimum elevation on that day, returning a positive (negative) value if the maximum record of the day comes after (before) the minimum record on the same day. We fit the FG distribution to the resultant data with $n = 366$ records using the frequentist method and the Bayesian method, with results presented in Table 3.2. The two inference methods produced very similar estimates for most parameters, although small differences were observed. For example, one would estimate the mode of daily maximum water elevation change to be -0.795 feet based on the frequentist method, but estimate it to be -0.486 feet using the Bayesian method. The discrepancy between these two mode estimates is minimal considering that the daily maximum water elevation changes range from -38 feet to 49.4 feet within this one-year period. In fact, taking into account the uncertainty in these point estimates, we do not interpret any of these

differences as statistically significant because a parameter estimate from one method always falls in the interval estimate for the same parameter from the other method according to Table 3.2. Using parameter estimates in Table 3.2 in the aforementioned R Shiny app, we obtained an estimated skewness of -0.102 and an estimated kurtosis of 6.384 based on the frequentist inference results, whereas the Bayes inference yielded an estimated skeweness of 0.058 and an estimated kurtosis of 6.074. Combining these two sets of results, we concluded that the underlying distribution of daily maximum water elevation change may be nearly symmetry, with outliers on both tails that cause tails heavier than that of a Gumbel distribution.

Table 3.2 Frequentist and Bayesian inferences about daily maximum water elevation changes of Lake Murray, South Carolina, United States. Besides parameter estimates (under point.est) and the estimated standard deviations of these parameter estimates (under $\widehat{\text{s.d.}}$), 95% confidence intervals of the parameters from the frequentist method, and 95% credible intervals from the Bayesian method are also provided (under lower 95 and upper 95).

	Frequentist				Bayesian			
parameter	point.est	$\widehat{\text{s.d.}}$	lower 95	upper 95	point.est	$\widehat{\text{s.d.}}$	lower 95	upper 95
θ	-0.795	0.796	-2.355	0.764	-0.486	0.694	-1.679	0.973
σ_1	5.186	0.541	4.124	6.247	5.399	0.651	4.534	6.916
σ_2	6.237	1.735	2.836	9.638	5.734	1.031	4.390	8.029
w	0.698	0.169	0.367	1.029	0.630	0.141	0.329	0.847

Figure 3.2 presents the estimated density functions from these two methods, in contrast with the estimated density curve resulting from fitting the data to a two-component normal mixture, and a kernel density estimate using a Gaussian kernel with the bandwidth selected according to the method proposed by Sheather and Jones (1991). The last estimate is fully nonparametric and served as a benchmark against which the other three density estimates were assessed graphically. The kernel density estimate is more flexible at describing varying tail behaviors, but such flexibility comes at the cost of statistical efficiency and interpretability. With the wiggly tails evident in Figure 3.2 for this estimate, we suspected certain level of overfitting of the kernel

density estimate. This often happens to kernel-based estimation of a function around a region where data are scarce, with a bandwidth not large enough for the region. Between the two FG density estimates, the difference is almost negligible. They both track the kernel density estimate closely over a wide range of the support around the mode. The mode of the estimated normal mixture density is close to the other three mode estimates, but the tails are much lighter than those of the other three estimated densities.

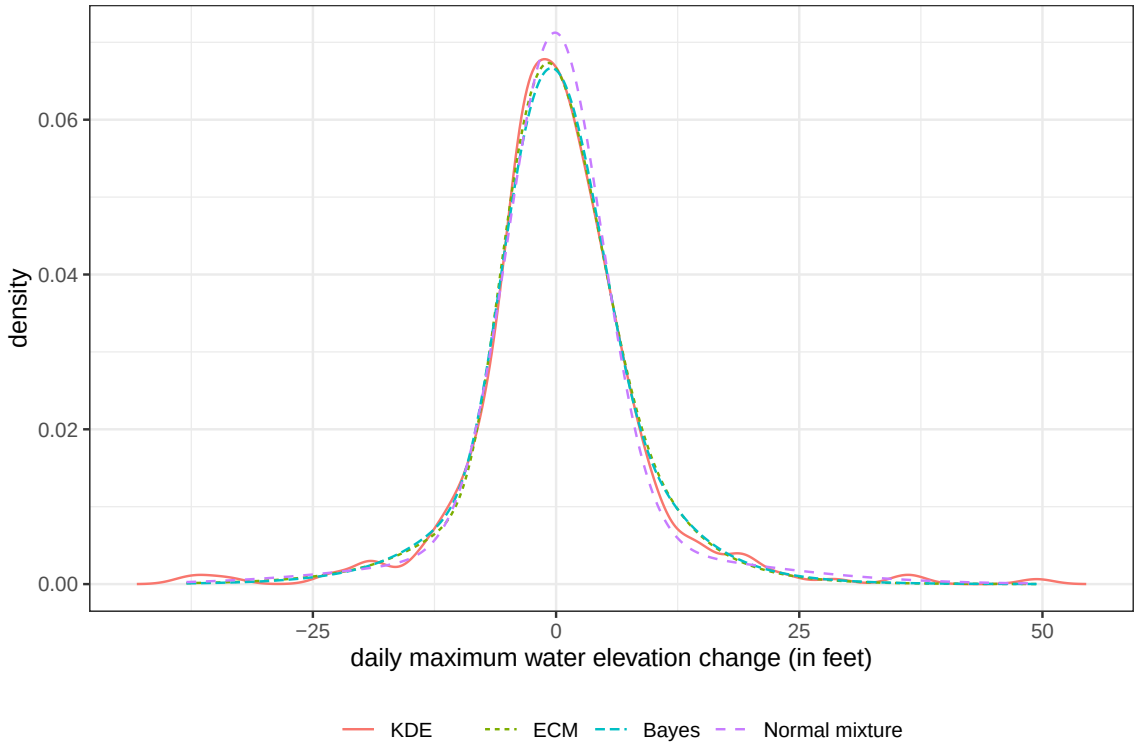


Figure 3.2 Four density estimates based on daily maximum water elevation changes in Lake Murray, including the kernel density estimate (solid line), the estimated FG density from the ECM algorithm (dotted line), the estimated FG density from the Bayesian method (dashed line), and the estimated normal mixture density (dash-dotted line).

Besides comparing the three parametric density estimates pictorially, we also used the Monte-Carlo based one-sample Kolmogorov–Smirnov test to assess the goodness

of fit. The p-values from this test are 0.223, 0.312, and 0.106 for the frequentist FG density estimate, the Bayesian FG density estimate, and the estimated normal mixture density, respectively. Although none of the p-values are low enough to indicate lack of fit (at significance level 0.05 for example), the p-value associated with the normal mixture is much lower than those for FG. This provides quantitative evidence that an FG distribution fits the current data better than a normal mixture. It is also worth noting that the Kolmogorov–Smirnov test is known to have low power to detect deviations from a posited distribution that occur in the tails (Mason and Schuenemeyer, 1983). This may explain the above-0.05 p-value for the normal mixture fit of the data even though the tail of this posited distribution may be too thin for the current data.

We used Stan to implement the Bayesian inference for the Lake Murray data, and the code and posterior output are given in the Appendix. The output provided there indicates that our MCMC chain has converged (see the `Rhat` statistics). The JAGS code for fitting the FG distribution is also given in the Appendix.

3.6 An application in criminology

Table 3.3 Frequentist and Bayesian modal regression models based on the FG distribution fitted to the crime data. Besides parameter estimates (under point.est) and the estimated standard deviations of these parameter estimates (under $\widehat{\text{s.d.}}$), 95% confidence intervals of the parameters from the frequentist method, and 95% credible intervals from the Bayesian method are also provided (under lower 95 and upper 95).

	Frequentist				Bayesian			
parameter	point.est	$\widehat{\text{s.d.}}$	lower 95	upper 95	point.est	$\widehat{\text{s.d.}}$	lower 95	upper 95
β_1	-0.166	0.071	-0.306	-0.026	-0.162	0.078	-0.315	-0.008
β_2	0.217	0.110	0.001	0.433	0.231	0.123	-0.009	0.475
β_3	0.067	0.013	0.042	0.093	0.067	0.014	0.039	0.095

With the location parameter θ signified in the FG distribution as the mode, it is straightforward to formulate a modal regression model that explores the relationship

Table 3.4 Mean regression model based on the normal distribution fitted to the crime data. Besides parameter estimates (under point.est) and the estimated standard deviations of these parameter estimates (under $\widehat{\text{s.d.}}$), 95% confidence intervals of the parameters

parameter	point.est	$\widehat{\text{s.d.}}$	lower 95	upper 95
β_1	0.467	0.161	0.142	0.792
β_2	1.140	0.224	0.689	1.591
β_3	0.068	0.034	0.000	0.136

between the response variable and predictors. To demonstrate the formulation of a modal regression model based on the FG distribution, we analyze a data set from Agresti et al. (2021) in the area of criminology. This data set contains the percentage of college education, poverty percentage, metropolitan rate, and murder rate for the 50 states in the United States and the District of Columbia from year 2003. The poverty percentage is the percentage of the residents with income below the poverty level; the metropolitan rate is defined as the percentage of population living in the metropolitan area; and the murder rate is the annual number of murders per $100,000$ people in the population.

We fit the following modal regression model to investigate the association between the murder rate (Y) and the aforementioned demographic variables,

$$Y \mid \boldsymbol{\beta}, \sigma_1, \sigma_2 \sim \text{FG}(\beta_0 + \beta_1 \times \text{ college } + \beta_2 \times \text{ poverty } + \beta_3 \times \text{ metropolitan }, \sigma_1, \sigma_2, w),$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]^\top$ includes all regression coefficients. For the prior elicitation in Bayesian inference, we assume that $\beta_0, \ldots, \beta_3 \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 10^4)$ and use the same priors for σ_1, σ_2 and w as those in Section 3.3.2. As a more conventional regression analysis to compare with our modal regression, we also fit the mean regression model assuming mean-zero normal model error to the data.

Table 3.3 shows the inference results from the modal regression model, and Table 3.4 presents the inference results from the mean regression model. At 5% significance level, both frequentist and Bayesian modal regression analyses confirm that there

exists a **negative** association between the percentage of college education and the murder rate, as well as a positive association between the metropolitan rate and the murder rate. In contrast, according to the inferred mean regression model, there is a **positive** association between the percentage of college education and the murder rate. Such claimed positive association is intuitively difficult to justify and contradicts with many published results in criminology (Hjalmarsson and Lochner, 2012; Lochner, 2020).

The scatter plot of the data in Figure 3.3 can shed some light on why one reaches to such a drastically different conclusion on a covariate effect when mean regression is considered in place of modal regression. As shown in Figure 3.3, the exists an obvious outlier, District of Columbia (DC), in panels of the first row of the scatter plot matrix for instance. Mean regression reacts to this one extreme outlier by inflating the covariate effect associated with the percentage of college education in the inferred mean regression function. Thanks to the heavy-tailed feature of the FG distribution, modal regression based on this distribution is robust to outliers, which strives to capture data features suggested by majority of the data and is not distracted by the extreme outlier when inferring covariate effects in this application.

3.7 Discussion

The mode had been an overlooked location parameter in statistical inference until recently when the statistics community witnessed a revived interest in modal regression among statisticians (Chen, 2018; Chacón, 2020; Feng et al., 2020; Xu et al., 2020; Ullah et al., 2021; Wang and Li, 2021; Xiang and Yao, 2022b). Historically, statistical inference for the mode have been mostly developed under the nonparametric framework for reasons we point out in Section 3.1. Existing semiparametric methods for modal regression only introduce parametric ingredients in the regression function, i.e., the conditional mode of the response, with the mode-zero error distribution left in a

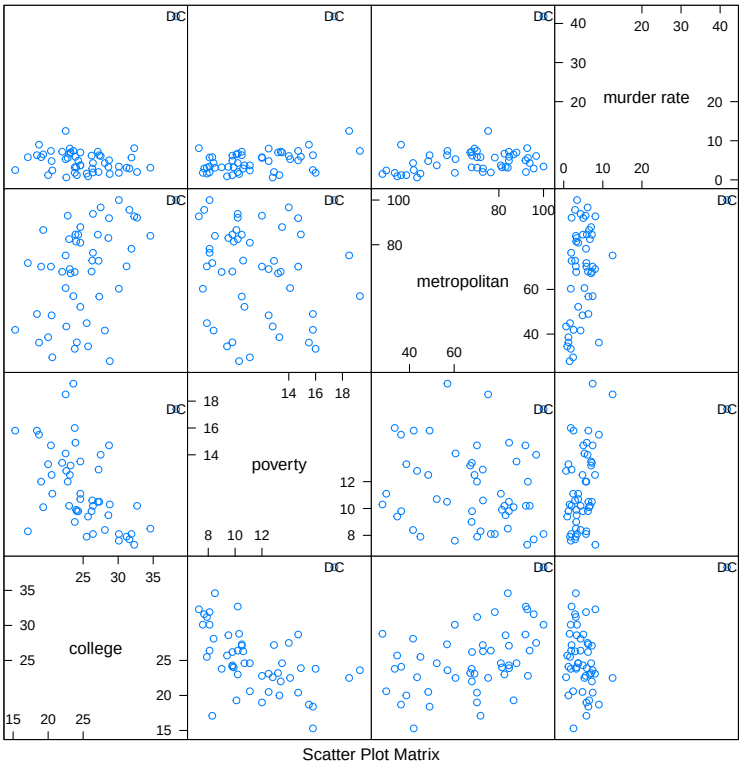


Figure 3.3 Scatter plot matrix of the crime data.

nonparametric form (Yao and Li, 2013; Liu et al., 2013; Zhang et al., 2013; Yang and Yang, 2014; Zhao et al., 2014; Krief, 2017; Tian et al., 2017; Li and Huang, 2019). The few recently proposed parametric modal regression models all impose stringent parametric assumptions on the error distribution (Bourguignon et al., 2020; Zhou and Huang, 2020, 2022). Our proposed flexible Gumbel distribution greatly alleviates concerns contributing to data scientists' reluctance to adopt a parametric framework when drawing inference for the mode. This new distribution is a heterogeneous mixture in the sense that the two components in the mixture belong to different Gumbel distribution families, which is a feature that shields it from the non-identifiability issue most traditional mixture distributions face, such as the normal mixtures. The proposed distribution is indexed by the mode along with shape and scale parameters, and thus is convenient to use to draw inference for the mode while remaining flexible. It is also especially suitable for modeling heavy-tailed data, whether the heaviness in

tails is due to extremely large or extremely small observations, or both. These are virtues of FG that cannot be achieved by the popular normal mixture and many other existing mixture distributions.

We develop the numerically efficient and stable ECM algorithm for frequentist inference for the FG distribution, and a reliable Bayesian inference method that can be easily implemented using free software, including Stan, JAGS, and BUGS. Compared with the more widely adopted mean regression framework, the modal regression model based on FG we entertained in Section 3.6 shows great potential in revealing meaningful covariate effects potentially masked by extreme outliers. With these advances made in this study, we open up new directions for parametric modal regression and semiparametric modal regression with a fully parametric yet flexible error distribution, and potentially nonparametric ingredients incorporated in the regression function.

Disclosure statement

Computer programs for implementing the FG distribution, related models and data used in this chapter are available at https://github.com/rh8liuqy/flexible_Gumbel.

Chapter 4

Bayesian Modal Regression based on Mixture Distributions

4.1 Introduction

Our study presented in earlier chapters and references therein suggest that modal regression models are useful additions to the well-established mean and median regression models. For unimodal and asymmetric distributions, intervals around the conditional mode typically have higher coverage probability than intervals of the same length around the conditional mean or median (Yao and Li, 2014; Xiang and Yao, 2022a). Consequently, prediction intervals from modal regression tend to be narrower than those for mean or median regression when data arise from a unimodal and skewed distribution. By construction, modal regression explores the relationship between the "most probable" value of Y given $\boldsymbol{X}$, and thus offers a highly interpretable representative value of the response. Thanks to the nature of the mode, modal regression is extremely robust to outliers that can obscure some inherent covariate effect suggested by the majority of observations, making it a worthy rival of median regression as an alternative to mean regression in regard to feature discovery.

A major challenge in building parametric modal regression models is constructing an appropriate distribution family that subsumes asymmetric, symmetric, light-tailed, and fat-tailed distributions. The flexible Gumbel distribution introduced in Chapter 3 contributes in this direction. In this chapter, we further propose the general unimodal distribution (GUD) family, which is a subfamily of the general two-component mixture

distribution family described in Section 4.3. Members of the GUD family have a location parameter as the mode, in addition to shape and scale parameters that control the skewness and tail behaviors. Thus, our framework is appropriate for both asymmetric *and* symmetric conditional distributions, as well as both light-tailed *and* fat-tailed distributions. In the extreme case, our framework can also model data from distributions without any finite moments, which we introduce in Section 4.3.3. We propose to estimate the conditional mode and the shape/scale parameters using a Bayesian approach. By placing appropriate prior distributions on model parameters, our modal regression models can be implemented straightforwardly using Markov chain Monte Carlo (MCMC) and provide natural uncertainty quantification through the posterior distributions.

4.1.1 Existing work on modal regression

Frequentist nonparametric modal regression has been the mainstream in the limited existing literature on modal regression (see Chen (2018) for a comprehensive review). The higher statistical efficiency and greater interpretability of covariate effects under a parametric framework motivate some recent development in frequentist parametric modal regression. For example, Aristodemou (2014) and Bourguignon et al. (2020) proposed a parametric modal regression model based on a gamma distribution for a positive response; Zhou et al. (2020) proposed two parametric modal regression models for a bounded response. Menezes et al. (2021) give a nice review on these and other parametric modal regression models for a bounded response. In contrast to these existing parametric modal regression models for *bounded* data, the modal regression models in the present manuscript are based on a *new* GUD family whose support is the *entire* real line. Furthermore, our work deals with *Bayesian* inference for modal regression.

The literature on Bayesian modal regression is even more sparse. Yu and Aris-

todemou (2012) proposed a nonparametric Bayesian modal regression model using Dirichlet process mixtures of uniform distributions. Zhou and Huang (2022) proposed a parametric Bayesian modal regression model based on a four-parameter beta distribution whose support is bounded yet unknown. Damien et al. (2017) introduced a more flexible parametric form of Bayesian modal regression using mixtures of triangular densities for a response with an unknown bounded support. Remaining in the parametric framework, a major strength of our proposed GUD family is that it naturally facilitates data-driven learning of the skewness and tails of the underlying distribution supported on the entire real line, while signifying the mode as the central tendency measure of the response.

4.1.2 Our contributions

The study presented in this chapter aims to widen the scope of Bayesian modal regression models and highlight the advantages of these models through analyses of datasets from real-life applications in several disciplines. We provide a unified framework of Bayesian modal regression models based on the GUD family that contains a large variety of unimodal distributions. We adopt the fully Bayesian approach via MCMC, so that inference of the conditional mode does not rely on asymptotic approximations. Our method is shown to provide reliable inference in small sample sizes. The fully Bayesian approach also comes with a convenient way of constructing prediction intervals from the posterior predictive distribution that is approximated using a simple random number generation algorithm for the GUD family. Indeed, the convenience in data generation via a data augmentation trick (to be discussed on Section 4.3) is yet another advantage of the GUD family. Finally, we exploit the model criterion known as the Bayesian leave-one-out expected log predictive density for model selection to help practitioners choose an appropriate distribution for their final analysis.

The main contributions of our work presented in this chapter can be summarized as follows:

1. We propose the GUD family that is suitable for Bayesian modal regression. The GUD family contains distributions that are symmetric or asymmetric, (non)normal, and/or fat-tailed.

2. We formulate rules of prior elicitation for the GUD family. In particular, we place a flat prior on regression coefficients, weakly informative priors on all other model parameters, and establish sufficient conditions under which the posterior distribution is proper.

3. We provide strategies for constructing prediction intervals and for selecting an appropriate likelihood for Bayesian modal regression analysis.

4. We illustrate the following benefits of our proposed Bayesian modal regression framework through simulation studies and data applications in economics, criminology, environmental science, and molecular biology: a) robustness to outliers, b) more precise prediction, and c) high interpratability of covariate effects.

The structure of the remainder of the chapter is as follows. In Section 4.2, we motivate our proposed Bayesian modal regression framework with two data applications. In Section 4.3, we formally define the GUD family and zoom in on several important members in the family. Section 4.4 introduces Bayesian inference for these modal regression models, including prior elicitation, posterior propriety, uncertainty quantification, and model selection. Section 4.5 provides simulation studies that illustrate the strengths of our methodology. In Section 4.6, we provide two additional data applications from environmental science and molecular biology. Section 4.7 concludes the chapter with some remarks about our Bayesian modal regression framework and several directions for future research.

4.2 Motivating applications

As a prelude to introducing our Bayesian modal regression framework, we first present results from applying the proposed methodology (to be elaborated in Sections 4.3 and 4.4) to datasets from the economics and the criminology literature.

4.2.1 Modeling highly right-skewed bank deposits

It is common knowledge to economists that wealth distributions are highly skewed to the right (Benhabib and Bisin, 2018). The cumulative nature of wealth not only has impact on individuals' net worth, but also has an influence on assets of large companies, including bank holding companies. In this example, we analyzed the deposits data of 50 banks and savings institutions in the United States on July 2, 2010 (Table 3.4.1 in Siegel (2012)).

Figure 4.1 presents the estimated density plot that results from fitting an intercept-only regression model based on the Double Two-Piece-Student-t, or DTP-Student-t, distribution (to be introduced in Section 4.3) to the dataset, along with the histogram of the underlying data. From this figure, we can see that the estimated mode using the DTP-Student-t distribution is close to the nonparametric mode estimate based on the histogram. This similarity and the close resemblance of the fitted density to the shape of the histogram indicate that the DTP-Student-t distribution is an adequate choice for the bank deposits data.

The other two measures of central tendency, i.e. the sample mean and median, are both shown to be larger than the estimated parametric mode in Figure 4.1. The sample mean, which equals 92.6 billion dollars, is obviously not a good measure of central tendency for most large banks and savings institutions in the United States. In particular, 40 of the 50 banks and savings institutions in our dataset had deposits *less* than 92.6 billion dollars on July 2, 2010. The sample median for this data is 40.5 billion dollars, indicating that 50% of banks in the dataset had deposits larger

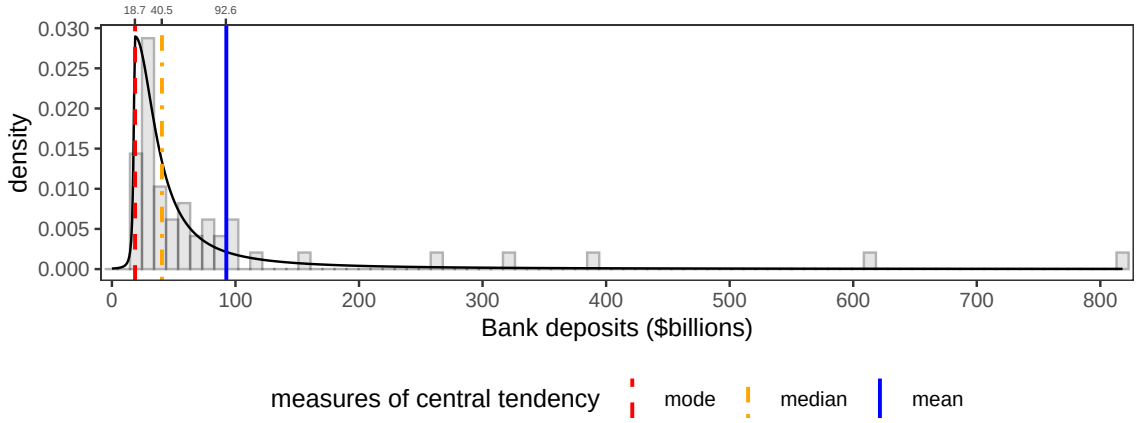


Figure 4.1 Deposits (in billions of dollars) of 50 banks and savings institutions in the United States on July 2, 2010. The solid black curve is the estimated density of the DTP-Student-t distribution. The three vertical lines mark locations of the sample mean (blue solid line), the sample median (orange dot-dashed line), and the estimated mode (red dashed line), respectively.

than 40.5 billion dollars while the other half had deposits smaller than 40.5 billion dollars. In spite of its high interpretability, the (sample) median is usually difficult to visualize either from a density plot or a histogram. In contrast, it is much easier for data analysts to locate and interpret the mode than the mean or median in Figure 4.1. The estimated mode using the DTP-Student-t distribution is where the density plot reaches its peak, and is close to where the histogram reaches to its peak. More specifically, the posterior mean of the mode is around 20 billion dollars, suggesting that banks in the United States are *most likely* to have deposits of around 20 billion dollars during that time.

4.2.2 Modal versus mean and median regression for analyzing murder rates

As a second motivating example, we analyze a dataset from Agresti et al. (2021) containing the murder rate, percentage of college education, poverty percentage, and metropolitan rate for the 50 states in the United States and the District of Columbia (D.C.) from 2003. The murder rate is defined as the annual number of murders

per 100,000 people in the population. The poverty percentage is the percentage of residents with income below the poverty level, and the metropolitan rate is defined as the percentage of population living in the metropolitan area.

At the stage of exploratory data analysis, we plotted the conditional scatter plot matrix of the U.S. crime data in Figure 4.2. From the first row of the conditional scatter plot matrix, a positive association between the poverty percentage and the murder rate, and also a positive association between the metropolitan rate and the murder rate are relatively evident; but an association between the college percentage and the murder rate is harder to perceive in the figure. Figure 4.2 also brings up a clear outlier, which is D.C. *Without* this outlier, there appears to exist a negative association between the college percentage and the murder rate.

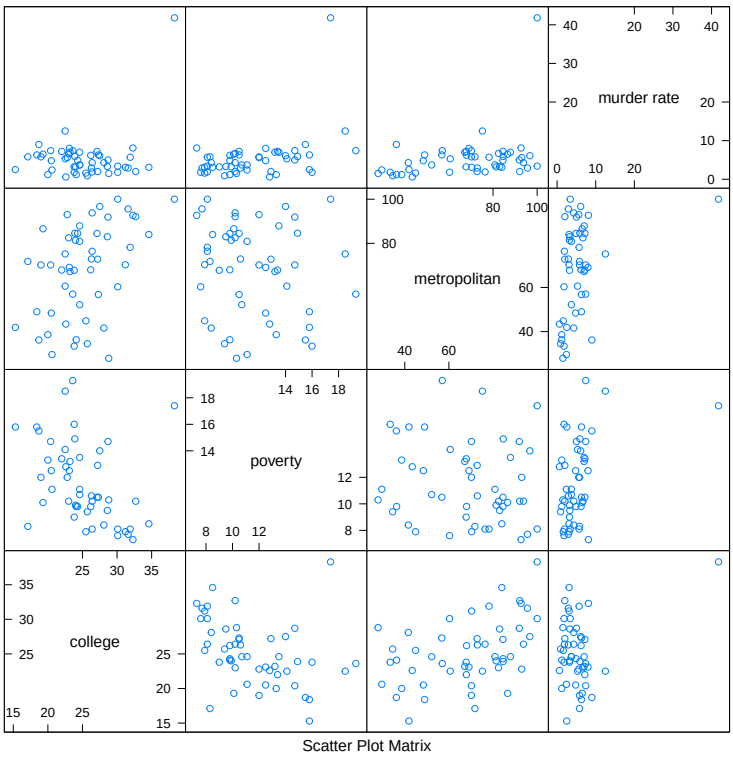


Figure 4.2 The conditional scatter plot matrices of the U.S. crime data.

To formally investigate the association between the murder rate (Y) and the

aforementioned variables, we fit the following models to the U.S. crime dataset:

$$\mathbb{M}(Y \mid \boldsymbol{\beta}) = \beta_0 + \beta_1 \times \text{college} + \beta_2 \times \text{poverty} + \beta_3 \times \text{metropolitan},$$

where $\mathbb{M}(\cdot)$ generically refers to the conditional mean, median, or mode. Table 4.1 presents the inference results from mean/median/modal regression models. All three models share some conclusions in common. Namely, all models determined that there were positive associations between the poverty percentage and the murder rate, and between the metropolitan rate and the murder rate. However, with a posterior credible interval (CI) of $(0.20, 0.74)$, the mean regression model (specified by (4.5.1)) implies that there exists a *positive* association between the college percentage and the crime rate, conditionally on the other covariates in the model. We believe that this inference result is difficult to justify, in light of existing results from the criminology literature that conclude a negative association between higher education attainment and crime (Lochner, 2020; Hjalmarsson and Lochner, 2012). On the other hand, with a CI of $(-0.27, 0.05)$, the Bayesian median regression model (formulated in (4.5.2)) concludes that the college percentage is *not* significantly associated with the murder rate, conditionally on the other covariates. Our Bayesian modal regression model with the Two-Piece scale-Student-t, or TPSC-Student-t, distribution (to be introduced in Section 4.3) draws a different conclusion. With a CI of $(-0.33, -0.06)$, our Bayesian modal regression model concludes that there is a *negative* association between the college percentage and the murder rate, which is more consistent with findings from the criminology literature. Lastly, according to the model criterion referred to as the expected log predictive density (ELPD, introduced in Section 4.4.3) in Table 4.1, the modal regression model based on the TPSC-Student-t likelihood yields the highest value of ELPD, indicating a better fit for the data than the mean and median regression models.

We repeated the above analyses after removing the D.C. outlier from the data. Now the median and modal regression models do in fact suggest a negative association

Table 4.1 Estimates of covariate effects for the mean/median/modal regression models fit to the U.S. crime dataset. The mean, 5% quantile, and 95% quantile of the posterior distribution of each covariate effect are listed under Mean, q5, and q95, respectively. ELPD stands for expected log predictive density.

Regression model	ELPD	Parameter (covariate)	Mean	q5	q95
Mean regression	-162.59	β_1 (college)	0.47	0.20	0.74
		β_2 (poverty)	1.14	0.76	1.51
		β_3 (metropolitan)	0.07	0.01	0.12
Median regression	-133.12	β_1 (college)	-0.12	-0.27	0.05
		β_2 (poverty)	0.44	0.22	0.67
		β_3 (metropolitan)	0.06	0.03	0.08
Modal regression	-123.27	β_1 (college)	-0.20	-0.33	-0.06
		β_2 (poverty)	0.24	0.01	0.46
		β_3 (metropolitan)	0.06	0.04	0.09

between the college percentage and the murder rate, whereas the mean regression model insists on lack of significant association between them. This exercise demonstrates that modal regression based on the proposed GUD family can be even more robust to outliers than median regression, and has a stronger potential in drawing reliable inferences and unveiling important features of data even in the presence of extreme outliers.

4.3 The family of general unimodal distributions

Having motivated our Bayesian modal regression framework and demonstrated its benefits on two real-life applications in Section 4.2, we now formally introduce the GUD family for Bayesian modal regression.

The probability density function (pdf) of a member belonging to the GUD family is a mixture of two pdfs, f_1 and f_2, given by

$$f(y \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = w f_1(y \mid \theta, \boldsymbol{\xi}_1) + (1 - w) f_2(y \mid \theta, \boldsymbol{\xi}_2). \tag{4.3.1}$$

In the mixture pdf (4.3.1), $w \in [0,1]$ is the weight parameter, $\theta \in (-\infty, +\infty)$ is the mode as the only location parameter in (4.3.1), $\boldsymbol{\xi}_1$ consists of parameters other than the location parameter in the pdf $f_1(\cdot \mid \theta, \boldsymbol{\xi}_1)$, and $\boldsymbol{\xi}_2$ is defined similarly for $f_2(\cdot \mid \theta, \boldsymbol{\xi}_2)$. Clearly, the GUD family belongs to the more general two-component mixture distribution family. One feature of GUD that makes it stand out from the bigger family of two-component mixture distributions is that the two component distributions of GUD share the same location parameter θ as the mode, a feature that makes GUD especially suitable for modal regression. In contrast, a two-component normal mixture for instance, as a widely referenced member in the bigger family, can be multimodal, and it is non-trivial to impose constraints on two normal components to guarantee unimodality (Sitek, 2016). Even after formulating a unimodal normal mixture, its mode may not have an analytical form (Behboodian, 1970). Many other members in the more general two-component mixture distribution family have the same pitfalls.

Besides unimodality, we reiterate and complement the following three restrictions on (4.3.1) to make the GUD family suitable and convenient for modal regression:

(R1) The pdfs $f_1(\cdot \mid \theta, \boldsymbol{\xi}_1)$ and $f_2(\cdot \mid \theta, \boldsymbol{\xi}_2)$ are unimodal at θ.

(R2) The pdfs $f_1(\cdot \mid \theta, \boldsymbol{\xi}_1)$ and $f_2(\cdot \mid \theta, \boldsymbol{\xi}_2)$ are left-skewed and right-skewed respectively.

(R3) The mixture pdf $f(\cdot \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ in (4.3.1) is continuous in its domain.

Restriction (R1) is already implied earlier when we stress that the two components in (4.3.1) share the same location parameter θ as the finite mode. In the context of modal regression, (R1) ensures that one can easily link a linear predictor $\boldsymbol{X}^\top \boldsymbol{\beta}$ with the conditional mode of Y. Because modal regression adds more value to mean/median regression when data are skewed and contain outliers, we impose (R2) to make members in GUD exhibit a wide range of skewness and tail behaviors. This

second restriction also solves the notorious label switching problem that many other two-component mixture distributions suffer from, because $f_1(\cdot \mid \theta, \boldsymbol{\xi}_1)$ and $f_2(\cdot \mid \theta, \boldsymbol{\xi}_2)$ satisfying (R2) must come from different distribution families in some strict sense, as opposed to, say, both coming from the normal family. According to Theorem 1 of Teicher (1963), this guarantees identifiability of all parameters associated with GUD. Lastly, (R3) eliminates ill-constructed pdfs whose mode may occur at a jump discontinuity.

Henceforth, when a random variable Y follows a distribution in the GUD family, we state that $Y \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \sim \text{GUD}\,(w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$. Like for other two-component mixture distributions, one may view $Y = ZX_1 + (1 - Z)X_2$, where $X_1 \mid \theta, \boldsymbol{\xi}_1 \sim f_1(\cdot \mid \theta, \boldsymbol{\xi}_1)$, $X_2 \mid \theta, \boldsymbol{\xi}_2 \sim f_2(\cdot \mid \theta, \boldsymbol{\xi}_2)$, and $Z \mid w \sim \text{Bernoulli}\,(w)$, with Z, X_1, and X_2 independent. This viewpoint gives rise to a data augmentation method outlined below for generating data from a GUD effortlessly:

(i) Sample $X_1 \mid \theta, \boldsymbol{\xi}_1 \sim f_1\,(\cdot \mid \theta, \boldsymbol{\xi}_1)$.

(ii) Sample $X_2 \mid \theta, \boldsymbol{\xi}_2 \sim f_2\,(\cdot \mid \theta, \boldsymbol{\xi}_2)$.

(iii) Sample $Z \mid w \sim \text{Bernoulli}(w)$.

(iv) $Y \leftarrow ZX_1 + (1 - Z)X_2$.

Having an efficient random number generation method is especially beneficial in constructing Bayesian prediction intervals, since the most common way to approximate the posterior predictive density is by drawing samples from the posterior predictive distribution during the MCMC iterations. We will continue our discussion about the Bayesian prediction intervals in Section 4.4.3.

Relating to existing literature, the GUD family *subsumes* several previously proposed distributions, such as those introduced in Fernández and Steel (1998) and Rubio and Steel (2015), as special cases. In what follows, we detail several examples of distributions from the GUD family.

4.3.1 The flexible Gumbel distribution

For predicting extreme events, the Gumbel distribution is a popular choice in many fields such as hydrology, earthquake forecasting, and insurance (Smith, 2003; Vidal, 2014; Shin et al., 2015a). The pdf of a Gumbel distribution for the maximum is

$$f_{\text{Gumbel}}\left(y \mid \theta, \sigma\right) = \frac{1}{\sigma} \exp\left\{-\frac{y-\theta}{\sigma} - \exp\left(-\frac{y-\theta}{\sigma}\right)\right\} \mathbb{I}\left(-\infty < y < \infty\right),$$

where $\theta \in \mathbb{R}$ is the mode as the location parameter, $\sigma > 0$ is the scale parameter, and $\mathbb{I}(\cdot)$ is the indicator function. To describe data that contains a mix of extremely large and extremely small events, Liu et al. (2022) proposed the flexible Gumbel (FG) distribution specified by the pdf

$$f_{\text{FG}}\left(y \mid w, \theta, \sigma_1, \sigma_2\right) = w f_{\text{Gumbel}}\left(-y \mid -\theta, \sigma_1\right) + (1-w) f_{\text{Gumbel}}\left(y \mid \theta, \sigma_2\right). \quad (4.3.2)$$

By mapping to (4.3.1), we have $f_1(y|\theta, \boldsymbol{\xi}_1) = f_{\text{Gumbel}}\left(-y \mid -\theta, \sigma_1\right)$ as the pdf of the left-skewed Gumbel distribution for the minimum. Similarly, we have $f_2(y|\theta, \boldsymbol{\xi}_2) = f_{\text{Gumbel}}\left(y \mid \theta, \sigma_2\right)$ as the pdf of the right-skewed Gumbel distribution for the maximum. We illustrate Bayesian modal regression based on the FG likelihood in Section 4.6.2. The FG distribution serves as a good choice of likelihood if the data is a mixture of extreme events, such as monthly maximum/minimum water elevation changes, and weekly heaviest/lightest traffic on a highway.

4.3.2 The double two-piece distribution

Rubio and Steel (2015) defined the Double Two-Piece (DTP) distribution by mixing two truncated distributions. For a pdf belonging to some location-scale family of the form $(1/\sigma) f\left(\left(y-\theta\right)/\sigma \mid \delta\right)$ that is unimodal at θ, with a scale parameter $\sigma > 0$ and a shape parameter δ, the pdf of the corresponding left θ-truncated distribution is

$$f_{\text{LT}}(y \mid \theta, \sigma, \delta) = \frac{2}{\sigma} f\left(\frac{y-\theta}{\sigma}\middle| \delta\right) \mathbb{I}(y < \theta), \quad (4.3.3)$$

and the corresponding right θ-truncated distribution is specified by the following pdf,

$$f_{\mathrm{RT}}(y \mid \theta, \sigma, \delta) = \frac{2}{\sigma} f\left(\frac{y-\theta}{\sigma}\middle| \delta\right) \mathbb{I}(y \geq \theta). \tag{4.3.4}$$

By mixing the pdfs in (4.3.3)-(4.3.4), we obtain the DTP pdf as

$$f_{\mathrm{DTP}}(y \mid \theta, \sigma_1, \sigma_2, \delta_1, \delta_2) = w f_{\mathrm{LT}}(y \mid \theta, \sigma_1, \delta_1) + (1-w) f_{\mathrm{RT}}(y \mid \theta, \sigma_2, \delta_2), \tag{4.3.5}$$

where

$$w = \frac{\sigma_1 f\left(0 \mid \delta_2\right)}{\sigma_1 f\left(0 \mid \delta_2\right) + \sigma_2 f\left(0 \mid \delta_1\right)} \tag{4.3.6}$$

as the weight chosen to produce a mixture distribution that satisfies (R3). Restrictions (R1) and (R2) are trivially satisfied by the construction of the left/right θ-truncated pdfs in (4.3.3)–(4.3.4). Thus, DTP distributions belong to the GUD family. Note, however, that our general GUD family (4.3.1) does not *require* the two component densities to be truncated, as we demonstrated earlier with the FG distribution (4.3.2).

As a concrete example, we consider the location-scale family as the three-parameter Student's t distributions, i.e., the non-standardized Student's t distributions, with location parameter θ, scale parameter $\sigma > 0$, and continuous degree of freedom $\delta > 0$ (Geweke, 1993). Following (4.3.3) and (4.3.4), one has the corresponding left-skewed truncated three-parameter Student's t distribution and the right-skewed truncated three-parameter Student's t distribution, respectively. This leads to the distribution defined according to (4.3.5) and (4.3.6) that we call the DTP-Student-t distribution. The DTP distribution family contains numerous distributions, all of which are suitable for modal regression (see Rubio and Steel (2015) for more). In the sequel, we concentrate on the DTP-Student-t distribution as a special member of the DTP distribution.

4.3.3 The two-piece scale distribution

By setting $\delta_1 = \delta_2 = \delta$ in (4.3.5), one obtains the pdf of a subfamily of the DTP family proposed in Fernández and Steel (1998), referred to as the two-piece scale (TPSC)

distribution family,

$$f_{\text{TPSC}}\left(y \mid w, \theta, \sigma, \delta\right) = w f_{\text{LT}}\left(y \mid \theta, \sigma\sqrt{\frac{w}{1-w}}, \delta\right) + (1-w) f_{\text{RT}}\left(y \mid \theta, \sigma\sqrt{\frac{1-w}{w}}, \delta\right). \tag{4.3.7}$$

We point out that in Fernández and Steel (1998), a shape parameter $\gamma = w^{0.5}(1-w)^{-0.5}$ is used instead of the weight parameter w when formulating the mixture pdf. We adopt the parameterization in (4.3.7) because we find it more straightforward to elicit a noninformative prior for w than placing a noninformative prior on γ.

Similar to the construction of the DTP-Student-t distribution, we can construct the TPSC-Student-t distribution by choosing the two component distributions to be the left and right θ-truncated three-parameter Student's t distributions. When $w = 0.5$, the TPSC-Student-t distribution converges to a normal distribution with mean θ and standard deviation σ as $\delta \to \infty$; and it reduces to a Cauchy distribution with mode θ and scale parameter σ when $\delta = 1$. Hence, even as a special case of the DTP-Student-t distribution, the TPSC-Student-t distribution is flexible enough to describe normally distributed data and non-normal data with extreme outlier(s) from distributions that do not have any finite moments. Since the TPSC-Student-t has fewer parameters than the DTP-Student-t distribution, it is an adequate choice for small datasets. On the other hand, the DTP-Student-t distribution may be preferred when there is moderate sample size. Certainly, one can conduct several rounds of modal regression analysis assuming different unimodal distributions for the response, such as the FG, DTP-Student-t, and TPSC-Student-t distributions, and then select the most appropriate model using the model selection criteria that we introduce in Section 4.4.3. All of these models can be easily implemented using the code developed for this work.

4.3.4 Type I GUD and type II GUD subfamilies

As illustrated in the preceding three subsections, the GUD family is a *generalization* of several previously proposed unimodal two-component mixture distributions. Figure 4.3 presents pdfs of FG, DTP-Student-t, and TPSC-Student-t distributions with different parameter specifications, which encompass asymmetric, symmetric, fat-tailed, *and* thin-tailed densities. In particular, the first panel in Figure 4.3 presents the density plot of FG distribution with varying scale parameters of the right-skewed component. As σ_2 becomes larger, the tails of FG distribution (especially the right tail) become fatter. In the second panel, we show that the FG distribution is symmetric if $w = 0.5$ and $\sigma_1 = \sigma_2$. With the weight parameter w surpassing 0.5 further, the pdf of FG distribution puts more weight on the left-skewed part, therefore, becomes more left-skewed. In the third panel, as the the scale parameter of the left-skewed component σ_1 increases, the left tail of the DTP-Student-t distribution becomes fatter while the right tail changes little, leading to distributions that are more left-skewed. The fourth panel shows the drastic change in the shape of the TPSC-Student-t pdf as one varies the scale parameter σ shared by both mixture components. The last panel presents the subtle changes in the tail behavior of TPSC-Student-t distributions with different values for the degree of freedom δ that is shared by both mixture components.

The GUD family can be further categorized into two subfamilies. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ denote the domains of $f_1(\cdot \mid \theta, \boldsymbol{\xi}_1)$ and $f_2(\cdot \mid \theta, \boldsymbol{\xi}_2)$ in the GUD pdf (4.3.1) respectively. If $\mathcal{D}_1 \cap \mathcal{D}_2 \neq \varnothing$, we call the mixture distribution the *type I GUD*. The FG distribution is an example of type I GUD. In Chapter A.2, we present the construction of the lognormal mixture distribution (logNM), which is another example of type I GUD. On the other hand, if $\mathcal{D}_1 \cap \mathcal{D}_2 = \varnothing$, then we have the *type II GUD*. The DTP distributions and the asymmetric Laplace distribution (ALD) (Koenker and Machado, 1999) belong to this subfamily of type II GUD. Figure 4.4 presents the partition of the GUD family

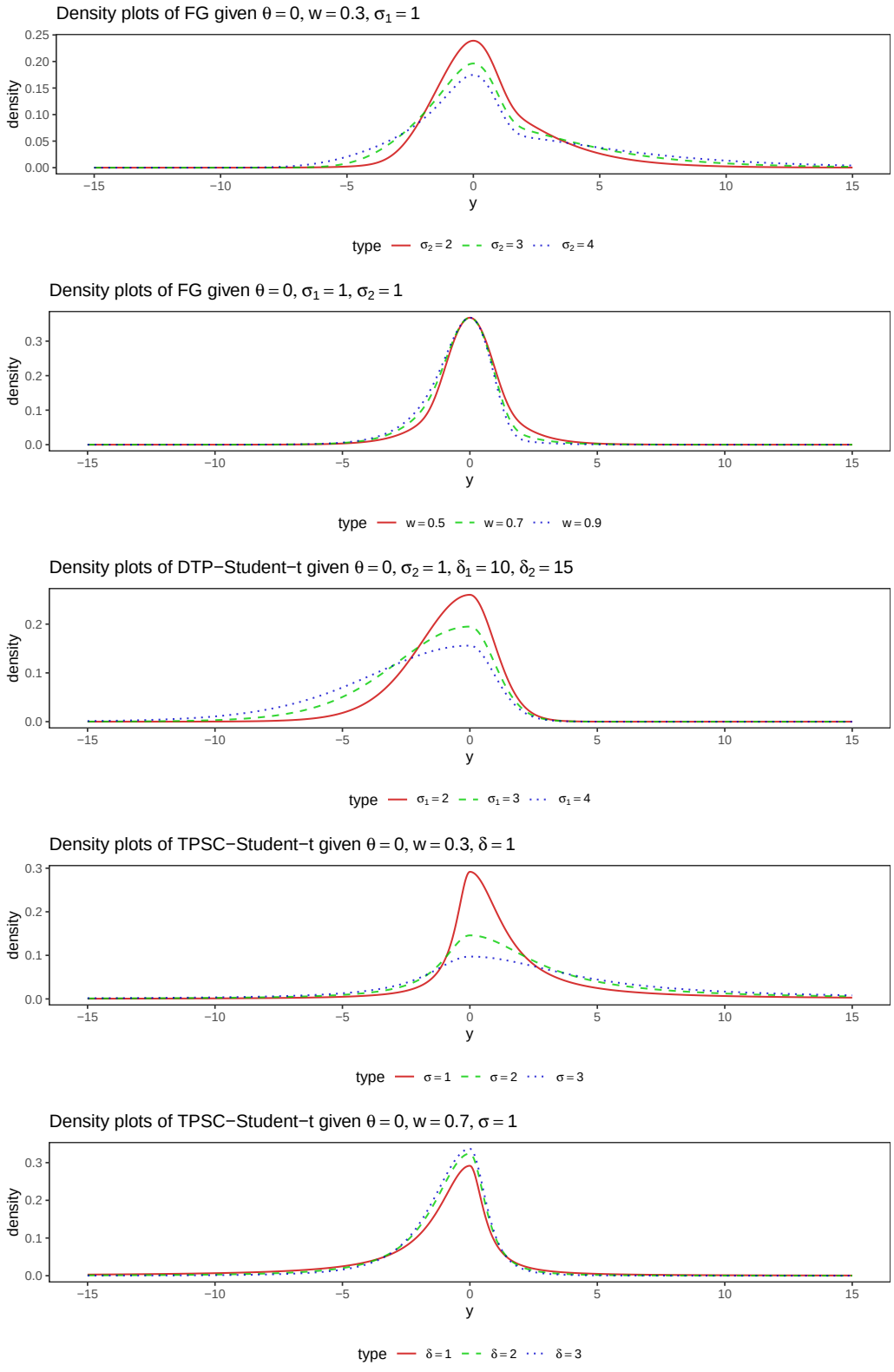


Figure 4.3 Density plots of different distributions in the GUD family with different parameter specifications.

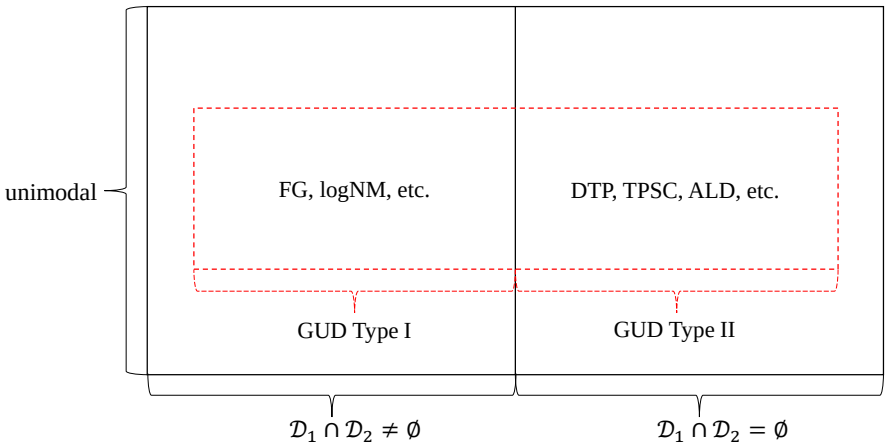


Figure 4.4 Venn diagram of the unimodal two-component mixture distributions.

into type I and type II GUD subfamilies.

4.4 Bayesian modal regression

Having defined the GUD family in Section 4.3, we are now in a position to introduce our Bayesian modal regression framework. In the remainder of the manuscript, we assume that we observe n independent pairs of observations $(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots, (\boldsymbol{X}_n, Y_n)$. Here, $\boldsymbol{X}_i := (X_{i1}, \ldots, X_{ip})^\top$ denotes a vector of p covariates for the ith observation. We let $\boldsymbol{X} := [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^\top$ denote an $n \times p$ design matrix with rows $\boldsymbol{X}_i^\top, i = 1, \ldots, n$. We assume exchangeability in the sense that, given $\boldsymbol{X}$ and all parameters, n observations in $\boldsymbol{Y} := (Y_1, \ldots, Y_n)$ are independent. Our goal is to conduct inference about the conditional *mode* of the response variable Y given the covariates $\boldsymbol{X}$.

4.4.1 Prior elicitation

For all modal linear regression models in this chapter, we assume that

$$Y_i \mid \boldsymbol{X}_i, w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \overset{\text{ind}}{\sim} \text{GUD}\left(w,\, \boldsymbol{X}_i^\top \boldsymbol{\beta},\, \boldsymbol{\xi}_1,\, \boldsymbol{\xi}_2\right), \text{ for } i = 1, \ldots, n, \tag{4.4.1}$$

where GUD generically refers to a member of the GUD family, and "ind" is the acronym for "independent." Recall that any member of the GUD family contains the

location parameter as its mode, which is $\boldsymbol{X}_i^\top \boldsymbol{\beta}$ as the conditional mode for Y_i given $\boldsymbol{X}_i$ in (4.4.1).

To conduct inference for our model in (4.4.1), we adopt a Bayesian approach where appropriate priors are placed on the model parameters $(w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$.We endow the weight parameter w with a noninformative Uniform$(0, 1)$ prior, and use weakly informative InverseGamma$(1, 1)$ or InverseGamma$(5, 5)$ priors for all positive parameters in $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$. As pointed out by Diebolt and Robert (1994), improper priors usually lead to improper posterior distributions for mixture distributions because of identifiability problems. Therefore, if $\boldsymbol{\xi}_1 \cap \boldsymbol{\xi}_2 = \varnothing$, then improper priors should *not* be used for $\boldsymbol{\xi}_1$ or $\boldsymbol{\xi}_2$.

On the other hand, a flat prior $p(\boldsymbol{\beta}) \propto 1$ on the regression coefficients $\boldsymbol{\beta}$ usually leads to a proper posterior distribution because both right and left skewed components share the same location parameter. In Section 4.4.2, we provide sufficient conditions under which a flat prior can be used for $\boldsymbol{\beta}$ such that the posterior distribution is proper. These sufficient conditions can be shown to hold for a variety of Bayesian modal regression models. All models going forward thus use a noninformative flat prior, $p(\boldsymbol{\beta}) \propto 1$, for $\boldsymbol{\beta}$.

Revisiting the three members of the GUD family discussed in Section 4.3, we have the Bayesian modal linear regression model based on the FG likelihood (4.3.2) formulated as follows,

$$\begin{aligned} Y_i \mid \boldsymbol{X}_i, w, \boldsymbol{\beta}, \sigma_1, \sigma_2 &\overset{\text{ind}}{\sim} \text{FG}\left(w,\, \boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma_1,\, \sigma_2\right), \text{ for } i = 1, \ldots, n, \\ w &\sim \text{Uniform}(0, 1), \\ \sigma_1, \sigma_2 &\overset{\text{i.i.d}}{\sim} \text{InverseGamma}(1, 1), \\ p(\boldsymbol{\beta}) &\propto 1, \end{aligned} \tag{4.4.2}$$

where "i.i.d" refers to "independent and identically distributed." Meanwhile, the Bayesian modal linear regression associated with the DTP-Student-t likelihood (4.3.5)

is specified by

$$\begin{aligned} Y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta}, \sigma_1, \sigma_2, \delta_1, \delta_2 &\overset{\text{ind}}{\sim} \text{DTP-Student-}t\left(\boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma_1,\, \sigma_2,\, \delta_1,\, \delta_2\right),\ \text{for } i = 1, \ldots, n, \\ \sigma_1, \sigma_2, \delta_1, \delta_2 &\overset{\text{i.i.d}}{\sim} \text{InverseGamma}(1, 1), \\ p(\boldsymbol{\beta}) &\propto 1. \end{aligned} \tag{4.4.3}$$

Recall that the weight parameter w of a DTP distribution is fully defined by its scale and shape parameters, so in this case, there is no need to choose a prior for w. Finally, the Bayesian modal linear regression associated with the TPSC-Student-t likelihood (4.3.7) is defined as

$$\begin{aligned} Y_i \mid \boldsymbol{X}_i, w, \boldsymbol{\beta}, \sigma, \delta &\overset{\text{ind}}{\sim} \text{TPSC-Student-}t\left(w,\, \boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma,\, \delta\right),\ \text{for } i = 1, \ldots, n, \\ w &\sim \text{Uniform}(0, 1), \\ \sigma, \delta &\overset{\text{i.i.d}}{\sim} \text{InverseGamma}(1, 1), \\ p(\boldsymbol{\beta}) &\propto 1. \end{aligned} \tag{4.4.4}$$

According to Proposition 4.4.2 in next subsection, all of the proposed Bayesian modal regression models (4.4.2)–(4.4.4) above have proper posterior distributions. Practitioners can construct various other Bayesian modal regression models using the same strategy shown above. In this chapter, we concentrate on the modal regression models based on the FG, DTP-Student-t, and TPSC-Student-t likelihoods for the sake of concreteness.

4.4.2 Sufficient conditions for posterior propriety

Since we use an improper prior, $p(\boldsymbol{\beta}) \propto 1$, for the regression coefficients $\boldsymbol{\beta}$ in our Bayesian modal regression models, it is important to check that the posterior distribution is proper. Theorem 4.4.1 gives sufficient conditions under which the GUD likelihood (4.3.1) with a flat prior on the mode/location parameter and suitably chosen priors on other model parameters lead to a proper posterior. Theorem 4.4.2 extends

this result to the regression setting. All the proofs for the theorems and propositions in this section can be found in Chapter A.1. We stress that our results are *nonasymptotic*; that is, our results apply for any *fixed* sample size n.

To ease the notation, let $f_Z(y \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) := f(y \mid w, \theta = 0, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ be the pdf of GUD family with the mode at 0. We can rewrite the pdf (4.3.1) as

$$f_Z(y - \theta \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = f(y \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2). \tag{4.4.5}$$

Theorem 4.4.1. Let $\Theta_{w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2}$ denote the parameter space of $w, \boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, with respective independent priors $p(w)$, $p(\boldsymbol{\xi}_1)$, and $p(\boldsymbol{\xi}_2)$. For any $n \geq 1$, if

$$\iiint_{\Theta_{w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2}} f_Z^{n-1}(0 \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)\, p(w) p(\boldsymbol{\xi}_1)\, p(\boldsymbol{\xi}_2)\, dw d\boldsymbol{\xi}_1 d\boldsymbol{\xi}_2 < \infty,$$

then the posterior distribution $p(w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \mid Y_1, \ldots, Y_n)$ is proper under a flat prior $p(\theta) \propto 1$.

Theorem 4.4.1 applies to the case where there is a single location parameter θ (as in the bank deposits application in Section 4.2.1). Next, we extend this result to the regression setting. Theorem 4.4.2 enables us to use the noninformative flat prior $p(\boldsymbol{\beta}) \propto 1$ for the regression coefficients $\boldsymbol{\beta}$ in Bayesian modal regression based on the GUD likelihood (4.3.1).

Theorem 4.4.2. Let $\boldsymbol{X}$ be a full rank design matrix with $p \leq n$ and finite entries. If

$$\iiint_{\Theta_{w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2}} f_Z^{n-p}(0 \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)\;\; p(w) p(\boldsymbol{\xi}_1)\, p(\boldsymbol{\xi}_2)\, dw d\boldsymbol{\xi}_1 d\boldsymbol{\xi}_2 < \infty,$$

then the posterior distribution $p(w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \mid \boldsymbol{X}, \boldsymbol{Y})$ is proper under a flat prior $p(\boldsymbol{\beta}) \propto 1$.

The sufficient conditions in Theorem 4.4.1 and 4.4.2 may seem abstract, and checking such conditions amounts to testing convergence of multiple integrals. The intuition behind these theorems is that, if the GUD likelihood with a mode of zero has a proper posterior distribution under suitably chosen priors on the scale/shape parameters, then the use of a flat prior $p(\boldsymbol{\beta}) \propto 1$ is acceptable.

Proposition 4.4.1. Suppose that $\boldsymbol{X}$ is full rank with $p \leq n$ and finite entries. Then the Bayesian modal regression models (4.4.2), (4.4.3), and (4.4.4) based on the FG, DTP-Student-t, and TPSC-Student-t likelihoods, respectively, have proper posterior distributions.

Proposition 4.4.1 confirms that under suitable regularity conditions on the design matrix $\boldsymbol{X}$, all of the regression models proposed in this chapter have proper posterior distributions. The proof of Proposition 4.4.1 relies on verifying the sufficient condition given in Theorem 4.4.2. Our proof provides a template for verifying posterior propriety for other Bayesian modal regression models (4.4.1) under the general GUD family.

Diebolt and Robert (1994) have argued that improper priors should in general not be used for Bayesian modeling of mixture distributions. We note, however, that the reasoning of Diebolt and Robert (1994) does not necessarily apply to the *location* parameter θ (or the mode). This is because the mode θ is shared by *both* left- and right-skewed components in our proposed GUD family of distributions. Therefore, we are able to derive sufficient conditions under which a totally noninformative flat prior $p(\theta) \propto 1$ or $p(\boldsymbol{\beta}) \propto 1$ can still be used to infer the conditional *mode*.

On the other hand, we recommend against using improper priors for any of the *non*-location parameters (i.e. the shape/scale parameters) in Bayesian modal regression based on the GUD family. We formalize this in Proposition 4.4.2 below. This proposition states that, for the GUD family, using an improper prior for *any* shape/scale parameter that is not shared by both components leads to an *improper* posterior distribution.

Proposition 4.4.2. If $\tau \in (\boldsymbol{\xi}_1 \cup \boldsymbol{\xi}_2) \setminus (\boldsymbol{\xi}_1 \cap \boldsymbol{\xi}_2)$, then using an improper prior for τ will lead to an improper posterior distribution.

In Chapter A.2, we provide a specific example of Proposition 4.4.2 for the logNM distribution (also introduced in the same section).

4.4.3 Uncertainty quantification and model selection

Letting $\boldsymbol{\Omega} = [w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2]^\top$, the posterior predictive distribution under our Bayesian modal regression model is defined as

$$\begin{aligned} p\left(Y_{\text{new}} \mid \boldsymbol{Y}, \boldsymbol{X}\right) &= \int_{\Theta} p\left(Y_{\text{new}} \mid \boldsymbol{\Omega}, \boldsymbol{Y}, \boldsymbol{X}\right) p\left(\boldsymbol{\Omega} \mid \boldsymbol{Y}, \boldsymbol{X}\right) d\boldsymbol{\Omega} \\ &= \int_{\Theta} p\left(Y_{\text{new}} \mid \boldsymbol{\Omega}, \boldsymbol{X}\right) p\left(\boldsymbol{\Omega} \mid \boldsymbol{Y}, \boldsymbol{X}\right) d\boldsymbol{\Omega}, \end{aligned} \tag{4.4.6}$$

where Θ denotes the parameter space, and the last line holds because of the conditional independence of Y_{new} and $\boldsymbol{Y}$. Obtaining an approximation of the posterior predictive (4.4.6) is computationally inexpensive. With the random number generation algorithm outlined in Section 4.3 for the GUD family, one can easily draw samples from $p(Y_{\text{new}} \mid \boldsymbol{\Omega}, \boldsymbol{X})$ during each iteration in our MCMC algorithm, and then obtain samples from the posterior predictive distribution $p\left(Y_{\text{new}} \mid \boldsymbol{Y}, \boldsymbol{X}\right)$. In this chapter, we use the `hdi` function in the `R` package `HDInterval` (R Core Team, 2022; Meredith et al., 2018), whose inputs are random samples generated from the posterior predictive distributions, to calculate the highest density intervals (HDI) with a pre-specified nominal level of coverage probability. We use 90% HDI intervals as the posterior prediction intervals for all mean/median/modal regression models that we consider in Sections 4.5 and 4.6.

Due to the inherent nature of the conditional mode, the HDI prediction intervals from modal regression models will usually be *narrower* than those constructed under mean or median regression models, while having the *same* amount of coverage (Yao and Li, 2014). From a statistical inference point of view, this is a very attractive property of our Bayesian modal regression models – we can obtain high coverage with tighter intervals. Prediction intervals from mean or median regression can sometimes be very conservative and contain many implausible values. We illustrate the benefits of more efficient inference from modal regression in Sections 4.5 and 4.6.

As mentioned in Section 4.4, there are many different GUD likelihoods that a practitioner can choose from in order to conduct Bayesian inference for modal

regression. We propose to use the Bayesian leave-one-out expected log posterior density as a model selection criterion for selecting the "best" GUD likelihood to use. The Bayesian leave-one-out expected log predictive density is defined as

$$\text{ELPD} = \sum_{i=1}^{n} \log p\left(Y_i \mid Y_{-i}\right), \tag{4.4.7}$$

where Y_{-i} represents all observations except the i-th observation. In (4.4.7), "ELPD" stands for the theoretical expected log predictive density. Intuitively, if a model fits the data well, its predicted value of Y_i given Y_{-i} should be close to the observed Y_i and $p\left(Y_i \mid Y_{-i}\right)$ should be large, for all $i = 1, \ldots, n$. Therefore, an adequate model tends to yield a high ELPD.

We apply the Pareto-smoothed importance sampling method (PSIS) of Vehtari et al. (2017) to obtain an estimate of ELPD. The PSIS estimation of ELPD has been implemented in an R package `loo`, which is compatible with the `Stan` programming language (Carpenter et al., 2017). We used the `Stan` programming language interfaced with R to implement all regression analysis in this chapter. When fitting multiple competing models to the same dataset, the model with the highest estimated ELPD is preferred. By a slight abuse of notation, we use ELPD to refer to the estimated ELPD in all empirical study presented in this chapter.

4.5 Simulation studies

We now present a few simulation studies which show that our Bayesian modal regression model is an excellent choice for modeling data that is heavily skewed. Under our simulation settings, simulated data was either left-skewed or right-skewed; and, in addition to the pronounced global conditional mode, there was also a small local mode. We compared our Bayesian modal regression models to Bayesian mean and median

regression models. The Bayesian mean regression used a normal likelihood, i.e.,

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \sigma, \boldsymbol{X}_i &\overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma^2), \text{ for } i = 1, \ldots, n, \\ \sigma &\sim \text{InverseGamma}(1, 1), \\ p(\boldsymbol{\beta}) &\propto 1. \end{aligned} \tag{4.5.1}$$

In line with the literature on parametric Bayesian quantile regression (Yu and Moyeed, 2001; Yu and Zhang, 2005), we also implemented Bayesian median regresson using the asymmetric Laplace distribution (ALD), with quantile parameter $p = 0.5$. That is, our Bayesian modal median regression model was

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \sigma, \boldsymbol{X}_i &\overset{\text{ind}}{\sim} \text{ALD}(\boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma,\, p = 0.5), \text{ for } i = 1, \ldots, n, \\ \sigma &\sim \text{InverseGamma}(5, 5), \\ p(\boldsymbol{\beta}) &\propto 1. \end{aligned} \tag{4.5.2}$$

We stress that in our simulation studies, *none* of the likelihoods used for mean, median, or modal regression was exactly the same as the data generating mechanism. Therefore, all considered regression models are "wrong," creating particularly realistic yet challenging scenarios under which we could more fairly compare the performance across these competing methods.

4.5.1 Left-skewed data

We generated $n = 30$ observations from the model,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where $\beta_0 = \beta_1 = 1$ and, for $i = 1, \ldots, 30$, $\epsilon_i \overset{\text{i.i.d}}{\sim} 0.05\mathcal{N}(-50, 1^2) + 0.95\mathcal{N}(0, 1^2)$ and $X_i \overset{\text{i.i.d}}{\sim} \text{Uniform}(0, 1)$. We then fit the mean/median/modal regression models to the simulated data. For modal regression, we fit the FG model (4.4.2), the DTP-Student-t model (4.4.3), and the TPSC-Student-t model (4.4.4). Among the modal regression models, we found that the TPSC-Student-t model had the highest ELPD. For the

sake of brevity, we present only the results from the models fit with the normal, ALD, and TPSC-Student-t likelihoods.

In Figure 4.5, we provide the empirical coverage rate and the average width of the posterior prediction intervals across $n = 30$ observations under each of mean/median/modal regression model. With the narrowest prediction interval for the same amount of coverage, results from the modal regression model clearly stand out in Figure 4.5. In addition, the modal regression model with the TPSC-Student-t likelihood had the largest ELPD. Therefore, it was the most appropriate model for the simulated data among the three candidate models in this replication.

We repeated this experiment 300 times. Table 4.2 shows the mean coverage rate, prediction interval width, and ELPD across the 300 replications. The mean regression model with the normal likelihood had the lowest average coverage rate and the widest posterior prediction intervals. Both the median and modal regression models had almost identical average coverage rate. However, the modal regression model had, on average, the narrowest prediction intervals. Since the modal regression model with the TPSC-Student-t likelihood had the largest average ELPD, we conclude that the modal regression model based on the TPSC-Student-t provided the best model fit.

4.5.2 Right-skewed data

In Section 4.5.1, we demonstrated the advantages of Bayesian modal regression models when the data was left-skewed. In this section, we investigate our model's ability to detect right-skewness. We followed the same simulation settings as those in Section 4.5.1, except the residual error was $\epsilon_i \overset{\text{i.i.d}}{\sim} 0.025\mathcal{N}(-25, 1^2) + 0.95\mathcal{N}(0, 1^2) + 0.025\mathcal{N}(50, 1^2)$ so that the simulated data was right-skewed, potentially with extremely large outliers and some outliers on the lower tail. We fit the mean/median/modal regression models to this simulated data.

Figure 4.6 shows that all three models achieved a coverage rate of 93% in one

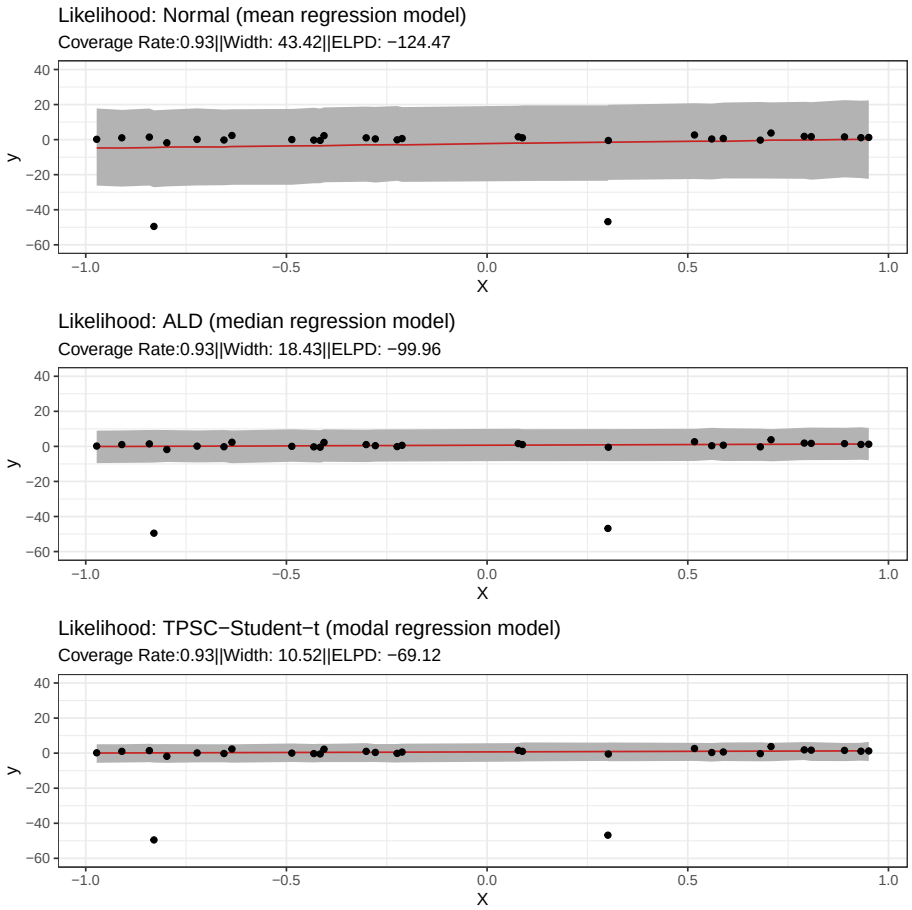


Figure 4.5 The gray shaded areas show the 90% posterior prediction intervals for the simulated left-skewed data. The solid red line is the estimated median from the posterior predictive distribution. The prediction intervals are narrower for Bayesian modal regression.

Table 4.2 Comparison of Bayesian mean, median, and modal regression models fitted to left-skewed data. Results were averaged across 300 Monte-Carlo replicates of left-skewed datasets. The empirical standard error associated with each Monte-Carlo average is provided in parenthesis following the average.

Likelihood (regression model)	Coverage Rate (%)	Width	ELPD
Normal (mean regression)	93.50 (0.19)	32.25 (1.08)	-104.76 (2.00)
ALD (median regression)	94.69 (0.22)	14.70 (0.49)	-84.81 (1.39)
TPSC-Student-t (modal regression)	**94.70** (0.22)	**8.36** (0.21)	**-**59.93 (0.64)

experiment. However, the modal regression model with the TPSC-Student-t likelihood (4.3.7) had the narrowest posterior prediction interval and the largest ELPD. Table 4.3 shows our results averaged across 300 repeated experiments. The modal regression model with the TPSC-Student-t likelihood had the highest average coverage rate, the narrowest posterior prediction intervals on average, and the largest average ELPD. Therefore, we conclude that the Bayesian modal regression model had the best performance in this right-skewed simulation study.

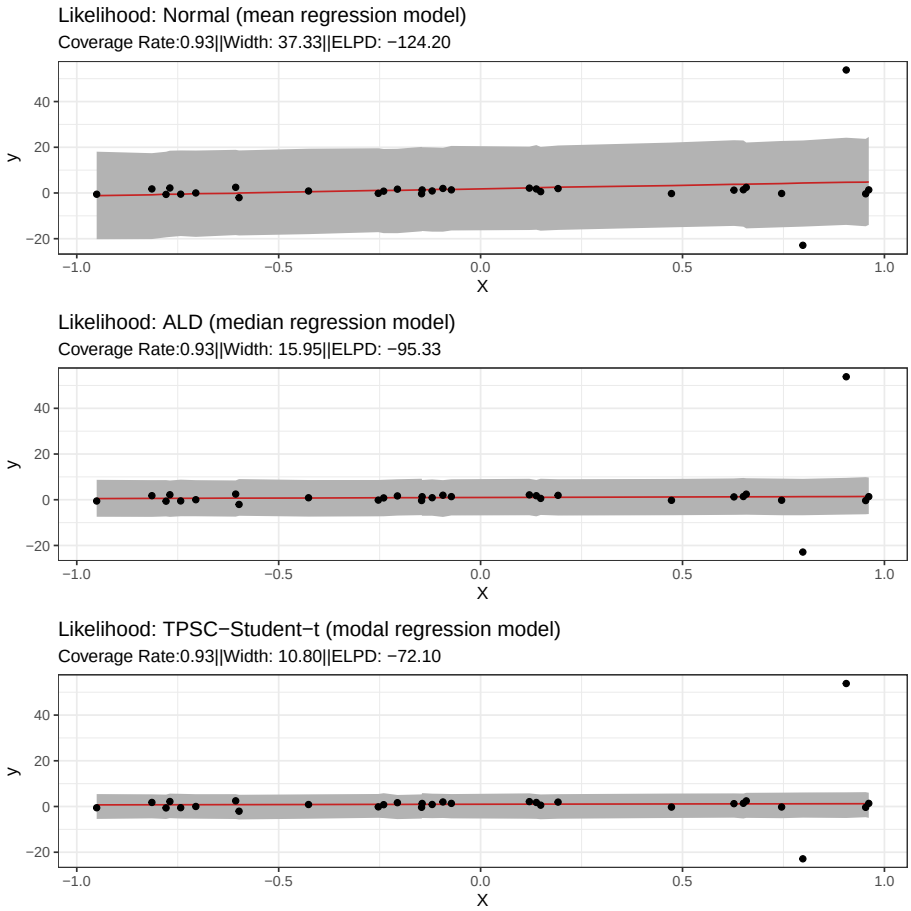


Figure 4.6 The gray shaded areas show the 90% posterior prediction intervals for the simulated right-skewed data. The solid red line is the estimated median from the posterior predictive distribution. The prediction intervals are narrower for Bayesian modal regression.

Table 4.3 Comparison of Bayesian mean, median, and modal regression models fitted to right-skewed data. Results were averaged across 300 Monte-Carlo replicates of right-skewed datasets. The empirical standard error associated with each Monte-Carlo average is provided in parenthesis following the average.

Likelihood (regression model)	Coverage Rate (%)	Width	ELPD
Normal (mean regression)	93.67 (0.18)	26.14 (0.95)	-99.41 (1.90)
ALD (median regression)	94.71 (0.22)	12.42 (0.40)	-79.25 (1.29)
TPSC-Student-t (modal regression)	**94.73** (0.22)	**8.20** (0.20)	**-**60.04 (0.65)

4.6 More data applications of Bayesian modal regression

4.6.1 Uncertainty quantification of air pollution

Inhalable particulate matter referred to as PM10 is any inhalable particle with a diameter of 10 micrometers and smaller. PM10 includes smoke, dust, and metals. In the environmental science literature, it has been found that low (resp. high) wind speeds are associated with high (resp. low) PM10 values (Cichowicz et al., 2020). In addition, many studies have found that PM10 is associated with heart disease, respiratory disease, and premature death (Schwartz, 1999; Zhao et al., 2017). We analyzed the `PM10` dataset from `http://lib.stat.cmu.edu/datasets/`. This dataset consists of air pollution information collected by the Norwegian Public Roads Administration at Alnabru in Oslo, Norway from October 2001 and August 2003. The response variable is the hourly measurements of the concentration of PM10, while the predictor is the hourly wind speed in meters per second. We fit the following mean/median/modal regression models to this dataset:

$$\mathbb{M}(Y \mid \boldsymbol{\beta}) = \beta_0 + \beta_1 \times \text{windspeed}\,.$$

Table 4.4 presents the parameter estimation results of three models. The mean regression model implies that the association between the PM10 concentration and the wind speed is not significant (CI of $(-2.64, 0.12)$). On the other hand, both the median and modal regression models capture the negative association between the

PM10 concentration and the wind speed (CIs of $(-2.77, -1.03)$ and $(-1.53, -0.54)$, respectively).

Although both median and modal regression detected the negative association between wind speed and PM10 concentration, we see from Figure 4.7 that the 90% prediction intervals for median regression (and mean regression) contain many negative values. As the minimum possible measure of PM10 concentration is zero, it is difficult to justify posterior prediction intervals for PM10 that contain many negative values. On the other hand, the 90% prediction intervals from the modal regression model only contain a tiny portion of negative values at very high wind speeds. This suggests that uncertainty quantification under Bayesian modal regression is more reliable and yields more practically meaningful results in this particular example.

Table 4.4 Parameter estimates obtained from the mean/median/modal regression models fitted to the air pollution data. The mean, 5% quantile, and 95% quantile of the posterior distribution of each regression coefficient are listed under Mean, q5, and q95, respectively.

Likelihood (regression model)	Parameter	Mean	q5	q95
Normal (mean regression)	β_0 (intercept)	41.94	36.80	47.10
	β_1 (windspeed)	-1.27	-2.64	0.12
ALD (median regression)	β_0 (intercept)	32.75	29.50	36.23
	β_1 (windspeed)	-1.87	-2.77	-1.03
TPSC-Student-t (modal regression)	β_0 (intercept)	9.67	7.47	11.75
	β_1 (windspeed)	-1.01	-1.53	-0.54

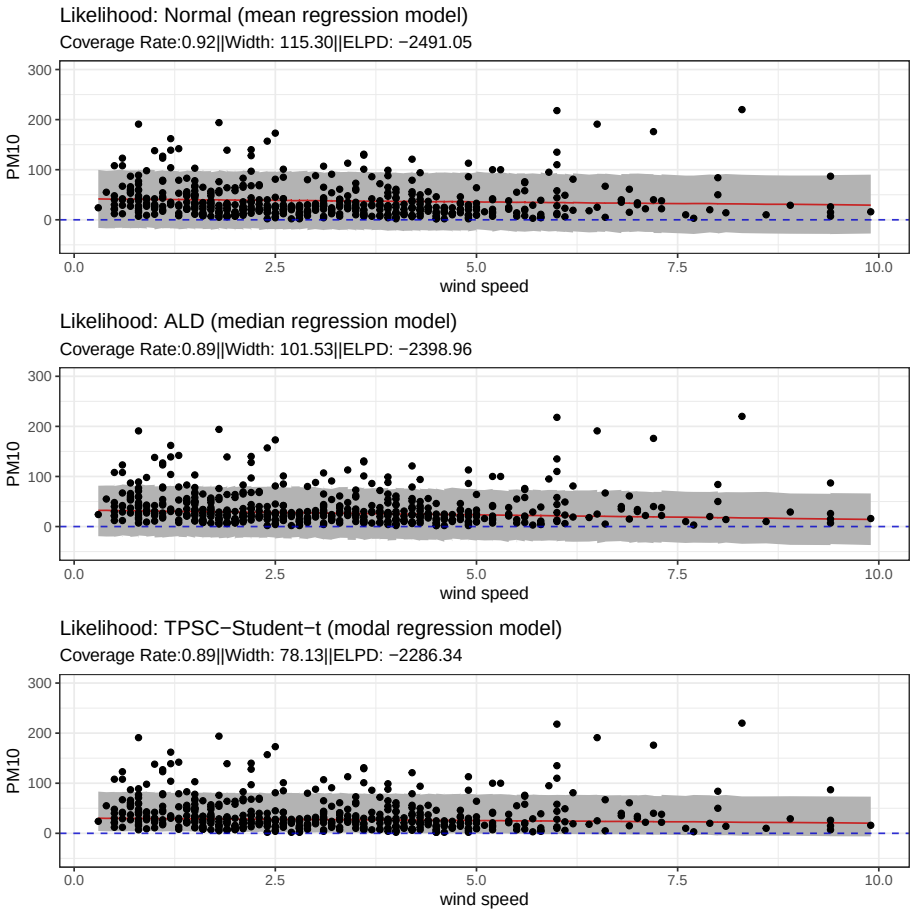


Figure 4.7 The shaded regions show the 90% posterior prediction intervals for the air pollution data. The blue dashed line is the reference line as the minimum possible value of PM10 (zero). The red solid line represents the estimated median from the posterior predictive distribution.

4.6.2 Detecting a quadratic relationship in serum data

Isaacs et al. (1983) analyzed the relationship between serum concentration (grams per litre) of immunoglobulin-G (IgG) in 298 children aged from 6 months to 6 years. IgG is an antibody that plays an important role in humoral and protective immunity (Van de Bovenkamp et al., 2016). There are ethical difficulties in taking repeated blood samples from healthy subjects. Therefore, researchers often use age as a proxy for determining the reference ranges for IgG in childhood. Previously, Yu and Moyeed (2001) analyzed serum data and modeled IgG concentration with a quadratic model in age. In the spirit of Yu and Moyeed (2001), we fit the following mean/median/modal

Table 4.5 Parameter estimates from the mean/median/modal regression models fitted to the serum data. The mean, 5% quantile, and 95% quantile of the posterior distribution of each regression coefficient are listed under Mean, q5, and q95, respectively.

Likelihood (regression model)	ELPD	Parameter	Mean	q5	q95
Normal (mean regression)	-627.13	β_0 (intercept)	3.08	2.45	3.72
		β_1 (Age)	0.97	0.45	1.49
		β_2 $\left(\text{Age}^2\right)$	-0.05	-0.13	0.04
ALD (median regression)	-638.16	β_0 (intercept)	2.82	2.12	3.57
		β_1 (Age)	1.12	0.53	1.68
		β_2 $\left(\text{Age}^2\right)$	-0.07	-0.16	0.03
FG (modal regression)	-623.15	β_0 (intercept)	2.37	1.84	2.89
		β_1 (Age)	1.16	0.72	1.59
		β_2 $\left(\text{Age}^2\right)$	-0.11	-0.18	-0.03

regression models to this dataset:

$$\mathbb{M}(Y \mid \boldsymbol{\beta}) = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Age}^2 \,.$$

Table 4.5 shows the parameter estimates from the models that we fit to this data. Based on the CIs for β_2, we see that only the modal regression model is able to detect the quadratic term (CI of $(-0.18, -0.03)$ for modal regression). This finding is somewhat consistent with Royston and Altman (1994) who concluded that a simple linear regression model was inadequate for this same dataset. Isaacs et al. (1983) also suggested that there was a quadratic relationship between the square root of IgG concentration and children's age.

Table 4.5 shows that the ELPD of the modal regression model based on the FG likelihood (4.4.2) is larger than the ELPD for both the mean or median regression models. In this example, the modal regression model not only provides a different viewpoint (i.e. that there exists a significant quadratic relationship between IgG and age), but it also fits the dataset better according to our model selection criterion.

4.7 Discussion

In this chapter, we have introduced a unifying Bayesian modal regression framework. Namely, we proposed a simple and flexible unimodal distribution family called the GUD family that is suitable for Bayesian modal regression. All members of the GUD family have a location parameter that is also the mode. Members of this family can be either symmetric or asymmetric, either thin-tailed or fat-tailed, depending on values of the shape and scale parameters.

Compared to mean and quantile regression, work on Bayesian modal regression analysis is quite scarce. Our work in this chapter aims to promote Bayesian modal regression as a complement to these other analyses. We demonstrated that our modeling framework based on the GUD family is very versatile and has wide applications in many fields such as economics (the bank deposit data in Section 4.2.1), criminology (the murder rate data in Section 4.2.2), environmental science (the air pollution data in Section 4.6.1), and molecular biology (the serum data in Section 4.6.2). In particular, we showed that Bayesian modal regression can reveal structures and detect potentially significant covariate effects that are missed by other Bayesian regression models.

To conduct Bayesian inference of the conditional mode, we provided prior elicitation procedures, along with the sufficient conditions under which a flat prior $p(\boldsymbol{\beta})$ on the regression coefficients $\boldsymbol{\beta}$ can be used. We proposed a method for constructing posterior prediction intervals and a model selection criterion based on the posterior predictive distribution. We demonstrated that our modal regression models provide very tight prediction intervals with high coverage, are robust to outliers, and have excellent interpretability. Our modal regression model framework is an especially appealing choice when the data is skewed and(or) contains (extreme) outliers.

The modal regression models considered here contain parametric assumptions, both about the data likelihood and the linear relationship between the covariates and the conditional mode. Instead of using the fully parametric models presented in

this chapter, one may prefer to use Bayesian semiparametric modal regression models instead. A Bayesian semiparametric modal regression model can be constructed either by modeling the conditional mode with a Gaussian process (i.e. we can relax the linearity assumption) and/or by replacing the GUD likelihood with a carefully constructed infinite mixture model that is indexed by the mode (i.e. we can relax the assumption of a known residual error distribution). These exciting extensions to Bayesian modal regression are the topics of ongoing work.

Chapter 5

The Gaussian Process Modal Regression

5.1 Introduction

A time continous stochastic process $\{f_t; t \in T\}$ is a Gaussian process if and only if the random vector $[f(t_1), \ldots, f(t_n)]^\top$ follows a multivariate normal distribution for all $t_1, \ldots, t_n \in T$ (Ross, 2014, Chapter 10). As pointed out in an excellent review paper by Li and Chen (2016), to apply Gaussian process in regression, we model a finite set of random function variables $\boldsymbol{f} = [f(\boldsymbol{X}_1), \ldots, f(\boldsymbol{X}_n)]^\top$ via a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and positive definitive variance-covariance $\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}$. The variance-covariance $\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}$ is generated by the kernel function $k(\cdot,\cdot)$ such that the entry in the i-th row and j-th column of $\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}$ is defined as

$$\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}(i,j) = k(\boldsymbol{X}_i, \boldsymbol{X}_j).$$

Common choices of kernel functions include the linear kernel function, the squared exponential kernel function, and the Ornstein-Uhlenbeck kernel function (for detailed definitions and other choices of kernel functions, see Duvenaud, 2014). In a Gaussian process regression (GPR) model, one often assumes that the response variable $\boldsymbol{y}$ given $\boldsymbol{f}$ follows a multivariate normal distribution with a diagonal variance-covariance matrix, in particular,

$$\boldsymbol{y} \mid \boldsymbol{f} \sim \mathcal{N}_n\left(\boldsymbol{f}, \sigma^2 \boldsymbol{I}_n\right), \tag{5.1.1}$$

where $\boldsymbol{I}_n$ stands for the identity matrix and σ^2 is the variance of noise. In the sequel, we refer the GPR model in (5.1.1) in conjunction with the assumption that $\boldsymbol{f} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}})$ as the mean GPR model. Wang and Shi (2014) extended the above

classic mean GPR model to allow $y_i \mid f(\boldsymbol{X}_i)$ follow a non-Gaussian distribution from an exponential family. Boukouvalas et al. (2012) proposed the quantile GPR model with an asymmetric Laplace distribution assumed for $y_i \mid f(\boldsymbol{X}_i)$ to signify the conditional quantile of the response. To the best of our knowledge, no existing GPR model has been developed based on the conditional mode of a response variable, i.e., letting $f(\boldsymbol{X_i})$ involve in formulating the mode of y_i given $\boldsymbol{X}_i$ and $\boldsymbol{f}$ follow a Gaussian process. Since the mean is well-known to be greatly influenced by outliers, as to be shown in Section 5.4, the performance of a mean GPR model can be unsatisfactory when the data contain extreme outliers and are highly skewed. As repeatedly demonstrated in previous chapters, the mode is resistant to outliers. This motivates our proposed modal GPR model for highly skewed data formulated in the next section.

5.2 Gaussian Process in Modal Regression

We define the modal Gaussian process regression model as, for $i = 1, \ldots, n$,

$$\begin{aligned} y_i \mid f\left(\boldsymbol{X}_i\right), \beta_0, \boldsymbol{\omega} &\sim \mathrm{GUD}\left(\theta_i = \beta_0 + f\left(\boldsymbol{X}_i\right), \boldsymbol{\omega}\right), \\ \boldsymbol{f} \mid \boldsymbol{\tau} &\sim \mathcal{N}_n(\mathbf{0}_n, \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}), \end{aligned} \tag{5.2.1}$$

where "GUD" refers to the general unimodal distribution family defined in Chapter 4 with the density function in (4.3.1), θ_i is the mode of y_i given $f\left(\boldsymbol{X}_i\right)$, β_0 is the intercept, $\boldsymbol{\omega} = [w, \boldsymbol{\xi}_1^\top, \boldsymbol{\xi}_2^\top]^\top$ represents weight parameter and scale/shape parameters in (4.3.1), and $\boldsymbol{\tau}$ are the parameters of the kernel function $k(\cdot, \cdot)$. For example, for the squared exponential kernel function, $k(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sigma_f^2 \exp\{-0.5\left(\boldsymbol{X}_i - \boldsymbol{X}_j\right)^2 / \ell^2\}$, we have $\boldsymbol{\tau} = [\sigma_f^2, \ell]^\top$. Certainly, any unimodal distribution family indexed by the mode (along with other parameters) as a location parameter can serve as the primary model in (5.2.1) when formulating a modal GPR model. For instance, the reparameterized Beta distribution introduced in Chapter 2 and the generalized biparabolic distribution proposed by García et al. (2009) are both viable options.

5.3 Statistical Inference

There are many widely applicable and well-accepted sampling methods incorporated in the Markov chain Monte Carlo (MCMC) algorithms for Bayesian learning based on GPR models. These include but not limited to the scaled Hamiltonian Monte Carlo and elliptical slice sampling (Bernardo et al., 1998; Murray et al., 2010). In recent years, the No-U-Turn Sampler (NUTS), proposed by Hoffman et al. (2014), has gained popularity in the machine learning community due to its self-tuning feature. Despite the excellent performance of NUTS, one major hurdle remains when applied to inferring a GPR model, which is the computational cost of the Cholesky decomposition of the variance-covariance matrix $\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}$. The Cholesky decomposition needs to be performed repeatedly for every MCMC iteration whenever the values of hyperparameters $\boldsymbol{\tau}$ change. We thus recommend using the exact Bayesian inference method via a probabilistic programming language, such as Stan, only when $n < 1000$. For larger data, we recommend that one employs scalable methods to infer a GPR model, some of which we introduce next.

To improve scalability without compromising predictive performance of a GPR model, numerous approximation methods have been proposed. Liu et al. (2020) provided a comprehensive review on scalable GPR models. Most recently, Riutort-Mayol et al. (2023) introduced a novel scalable GPR model based on a basis function approximation via Laplace eigenfunctions. This approach is specifically designed for stationary covariance functions and has shown promising results in terms of accuracy and efficiency. Moreover, it is easy to implement in probabilistic programming languages like Stan and is not limited to the Gaussian likelihood. We will apply this approach to implement scalable modal GPR modeling based on large data in our follow-up study. For now, we focus on demonstrating the potential of our proposed modal GPR models when the sample size is small or moderate so that an exact Bayesian inference method suffices to draw inference.

5.4 Simulation Study

We carry out a simulation study to demonstrate that, for data contain extreme outliers, the modal GPR model is more suitable than the mean GPR model.

For $i = 1, ..., n$, we generate data following three steps outlined below:

1. Simulate $X_i \sim \text{uniform}(-10, 10)$.

2. Simulate $\epsilon_i \sim \text{FG}(\text{mode} = 0, \sigma_1 = 100, \sigma_2 = 1, w = 0.8)$.

3. $y_i \leftarrow 0.2X_i \sin(X_i) + \epsilon_i$.

Given one simulated data set of size $n = 100$, we carried out exact Bayesian inference, first assuming a mean GPR model, then assuming our proposed modal GPR model. We then obtain the posterior mean of the mean regression function from the former inference procedure, and the posterior median of the modal regression function from the latter. These two posterior results are depicted in Figure 5.1. We use the posterior median to summarize posterior inference on the modal regression function because the resultant posterior distribution associated with the modal GPR model is highly skewed, making the posterior mean an inadequate summary of the posterior distribution of the mode function. This is unlike when one assumes a mean GPR model, where the resultant posterior distribution of the mean regression function remains a multivariate normal.

As evidenced in the zoom-in view of the posterior estimate of the modal function and that of the mean function in Figure 5.1 (see bottom panels), inference resulting from assuming a modal GPR model for this highly left-skewed data are able to provide a reliable estimate of the true mode function. Moreover, one achieves much tighter credible intervals for the target regression function when assuming a modal GPR model than when assuming a mean GPR model. Despite the existence of extreme outliers, the modal GPR model even captures the curvature of the true mode very

well. In contrast, the mean GPR model consistently underestimates the true mean regression function.

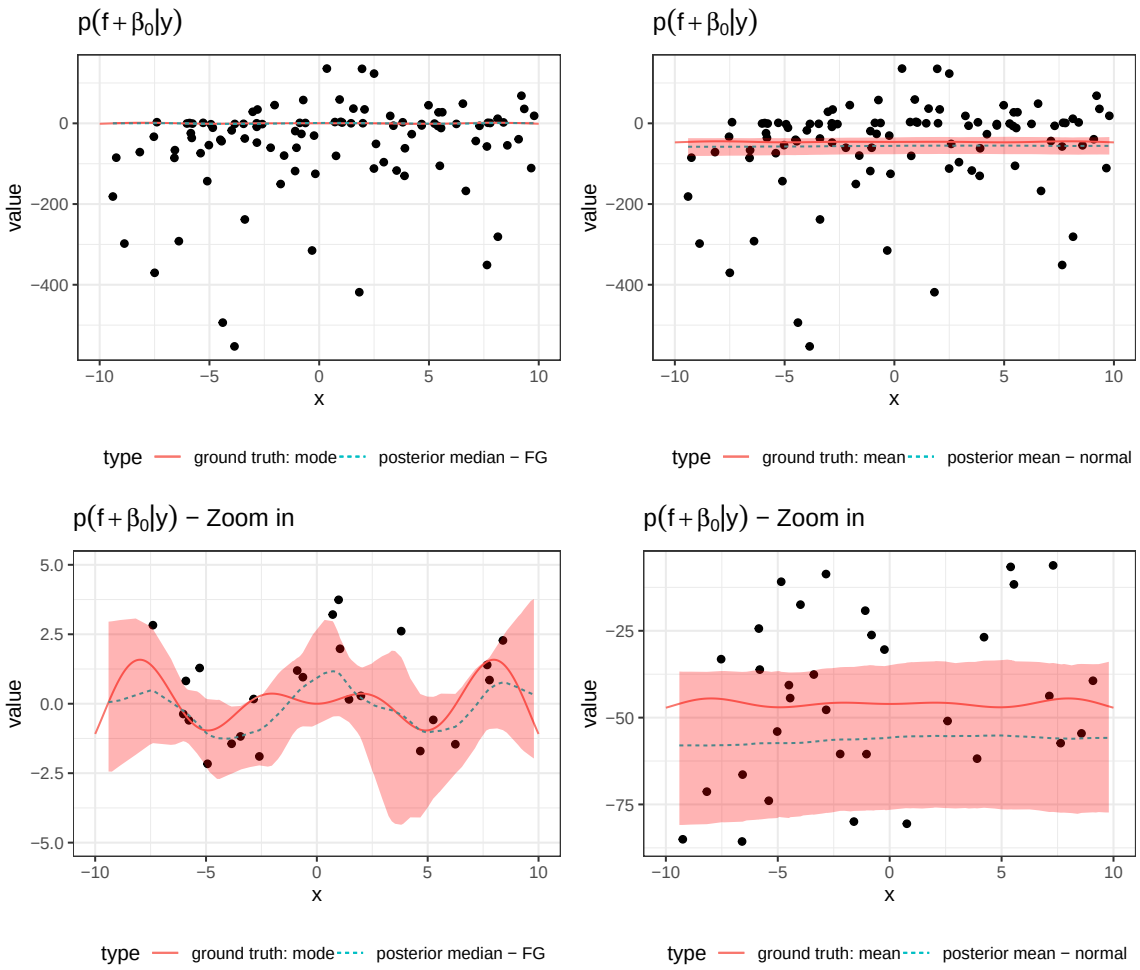


Figure 5.1 Comparison between the modal GPR model and the mean GPR model based on the same data set. The red solid lines in left/right column represents the true mode/mean of the simulated data respectively. The cyan dashed line in the left/right column represents the posterior median/mean from the modal GPR model and the mean GPR model respectively. The red bands in both columns represent 90% credible intervals.

5.5 Data Application

The Summer Olympic Games, often referred to as the Summer Olympics, is a major international multi-sport event that was typically held once every four years. The

first modern Olympic Games were held in 1896, and since then, the marathon, a long-distance foot race, has been one of the most prominent events of the Summer Olympics. Intrigued by the exceptional human performance in long-distance foot races, we acquired a dataset comprising the pace of Olympic male marathon champions from Athens 1896 to London 2012, measured in minutes per kilometer. Notably, an outlier is apparent in the year 1904. This outlier can be attributed to the poorly organized marathon held in St. Louis that year, which resulted in a pace equivalent to 5.222 minutes per kilometer.

Figure 5.2 displays the estimated central tendency (mode/mean), along with the corresponding 90% credible intervals, from the modal GPR model with the TPSC-Student-t distribution (4.3.7) assumed for the likelihood formulation, and the mean GPR model with a normal model error in the primary model. Again, in this application, the modal GPR model yields a narrower credible interval than the mean GPR model for the same coverage rate. Furthermore, the ELPD model selection criteria (4.4.7) indicates that the modal GPR model is superior to the mean GPR model for the given data.

5.6 Discussion

This study is ongoing, and we are currently conducting additional simulation studies to compare the performance of mean, median, and modal GPR models. We also plan to include real data application examples and evaluate the performance of scalable modal GPR models. An interesting avenue for future research would be to apply the modal GPR model to spatiotemporal analysis. While GPR models are capable of analyzing spatiotemporal data, most existing models are limited to mean GPR models, and we have not found any modal GPR models for spatiotemporal data. Previous work in spatiotemporal analysis using the mean GPR models includes studies by He et al. (2014), Hyun et al. (2016), and Sarkar et al. (2019).

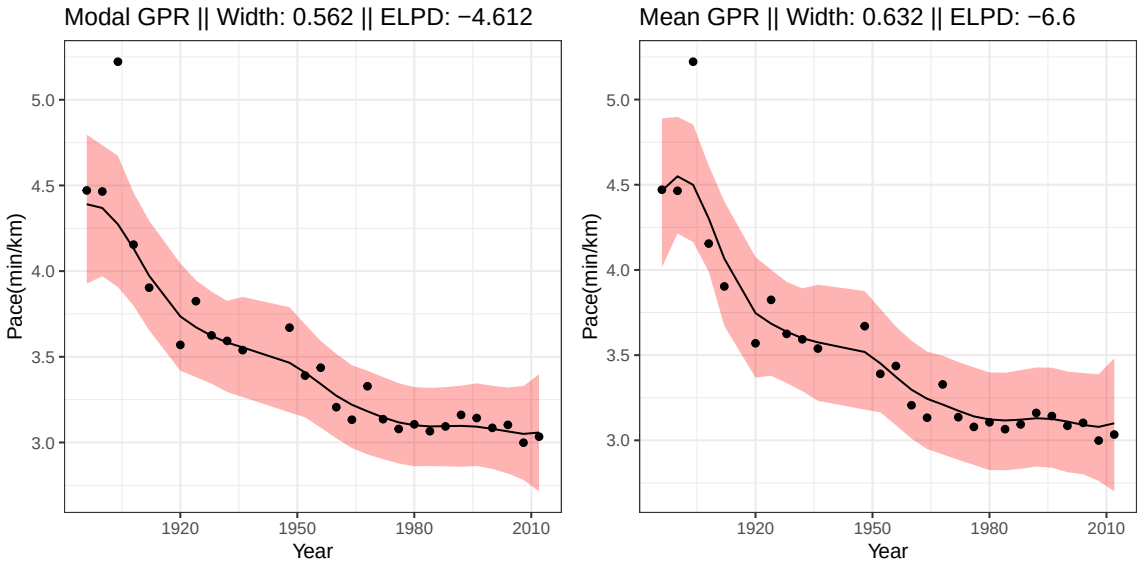


Figure 5.2 Comparison between the modal GPR model (based on the TPSC-Student-t distribution) and the mean GPR model (based on the normal distribution). The solid black lines depict the posterior median of $\boldsymbol{f} + \beta_0$, while the red bands show the 90% credible interval of $\boldsymbol{f} + \beta_0$.

Bibliography

Agresti, A., Franklin, C., and Klingenberg, B. (2021). *Statistics: The Art and Science of Learning from Data*. Pearson Education, 5 edition.

Aristodemou, K. (2014). *New regression methods for measures of central tendency*. PhD thesis, Brunel University.

Bali, T. G. (2003). An extreme value approach to estimating volatility and value at risk. *The Journal of Business*, 76(1):83–108.

Behboodian, J. (1970). On the modes of a mixture of two normal distributions. *Technometrics*, 12(1):131–139.

Benhabib, J. and Bisin, A. (2018). Skewed wealth distributions: Theory and empirics. *Journal of Economic Literature*, 56(4):1261–91.

Bernardo, J., Berger, J., Dawid, A., Smith, A., et al. (1998). Regression and classification using gaussian process priors. *Bayesian statistics*, 6:475.

Boos, D. D. and Stefanski, L. A. (2013). *Essential statistical inference: theory and methods*, volume 591. Springer.

Boukouvalas, A., Barillec, R., and Cornford, D. (2012). Gaussian process quantile regression using expectation propagation. *arXiv preprint arXiv:1206.6391*.

Bourguignon, M., Leão, J., and Gallardo, D. I. (2020). Parametric modal regression with varying precision. *Biometrical Journal*, 62(1):202–220.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Chapman and Hall/CRC.

Buonaccorsi, J. P., Romeo, G., and Thoresen, M. (2018). Model-based bootstrapping when correcting for measurement error with application to logistic regression. *Biometrics*, 74(1):135–144.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Carroll, R. J., Freedman, L., and Pee, D. (1997). Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error models. *Biometrics*, 53(4).

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC.

Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review*, 88(1):122–141.

Chen, Y.-C. (2018). Modal regression using kernel density estimation: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(4):e1431.

Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., and Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics*, 44(2):489–514.

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16(1):31–41.

Cichowicz, R., Wielgosiński, G., and Fetter, W. (2020). Effect of wind speed on the level of particulate matter pm10 concentration in atmospheric air during winter season in vicinity of large combustion plant. *Journal of Atmospheric Chemistry*, 77(1):35–48.

Cooray, K. (2010). Generalized Gumbel distribution. *Journal of Applied Statistics*, 37(1):171–179.

Dalenius, T. (1965). The mode–a neglected statistical parameter. *Journal of the Royal Statistical Society. Series A (General)*, 128(1):110–117.

Damien, P., Walker, S., et al. (2017). Bayesian mode regression using mixtures of triangular densities. *Journal of Econometrics*, 197(2):273–283.

Dawley, S., Zhang, Y., Liu, X., Jiang, P., Tick, G., Sun, H., Zheng, C., and Chen, L. (2019). Statistical analysis of extreme events in precipitation, stream discharge, and groundwater head fluctuation: distribution, memory, and correlation. *Water*, 11(4):707.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Desikan, R. S., Cabral, H. J., Settecase, F., Hess, C. P., Dillon, W. P., Glastonbury, C. M., Weiner, M. W., Schmansky, N. J., Salat, D. H., and Fischl, B. (2010). Automated mri measures predict progression to alzheimer's disease. *Neurobiology of Aging*, 31(8).

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56(2):363–375.

Duvenaud, D. (2014). *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge.

Feng, Y., Fan, J., and Suykens, J. (2020). A statistical learning approach to modal regression. *Journal of Machine Learning Research*, 21(2):1–35.

Fernández, C. and Steel, M. F. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Fuller, W. A. (2009). *Measurement Error Models*. John Wiley & Sons.

García, C. B. G., Pérez, J. G., and Rambaud, S. C. (2009). The generalized biparabolic distribution. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 17(03):377–396.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1).

Geweke, J. (1993). Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, 8(S1):S19–S40.

Gumbel, E. J. (1941). The return period of flood flows. *The Annals of Mathematical Statistics*, 12(2):163–190.

He, X. and Liang, H. (2000). Quantile regression estimates for a class of linear and partially linear errors-in-variables models. *Statistica Sinica*, 10(1):129–140.

Hjalmarsson, R. and Lochner, L. (2012). The impact of education on crime: International evidence. *CESifo DICE Report*, 10(2):49–55.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, 16(9):965–980.

Huang, X., Stefanski, L. A., and Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika*, 93(1):53–64.

Isaacs, D., Altman, D., Tidmarsh, C., Valman, H., and Webster, A. (1983). Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: Reference ranges for igg, iga, igm. *Journal of Clinical Pathology*, 36(10):1193–1196.

Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., Shaw, L. M., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G., Pankratz, V. S., Donohue, M. C., and Trojanowski, J. Q. (2013). Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207–216.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310.

Koutsoyiannis, D. (2004). Statistics of extremes and estimation of extreme rainfall: I. theoretical investigation / statistiques de valeurs extrêmes et estimation de

précipitations extrêmes: I. recherche théorique. *Hydrological Sciences Journal*, 49(4).

Krief, J. M. (2017). Semi-linear mode regression. *The Econometrics Journal*, 20(2):149–167.

Lee, M.-J. (1989). Mode regression. *Journal of Econometrics*, 42(3):337–349.

Lee, M.-J. (1993). Quadratic mode regression. *Journal of Econometrics*, 57(1-3):1–19.

Li, P. and Chen, S. (2016). A review on gaussian process latent variable models. *CAAI Transactions on Intelligence Technology*, 1(4):366–376.

Li, X. and Huang, X. (2019). Linear mode regression with covariate measurement error. *Canadian Journal of Statistics*, 47(2):262–280.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. (2020). When gaussian process meets big data: A review of scalable gps. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423.

Liu, J., Zhang, R., Zhao, W., and Lv, Y. (2013). A robust and efficient estimation method for single index models. *Journal of Multivariate Analysis*, 122:226–238.

Liu, Q., Huang, X., and Zhou, H. (2022). The flexible gumbel distribution: A new model for inference about the mode. *arXiv preprint arXiv:2212.01832*.

Loaiciga, H. A. and Leipnik, R. B. (1999). Analysis of extreme hydrologic events with gumbel distributions: marginal and additive cases. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 13(4):251–259.

Lochner, L. (2020). Education and crime. In *The Economics of Education*, pages 109–117. Elsevier.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067.

Maclaren, J., Han, Z., Vos, S. B., Fischbein, N., and Bammer, R. (2014). Reliability of brain volume measurements: A test-retest dataset. *Scientific Data*, 1(1).

Mason, D. M. and Schuenemeyer, J. H. (1983). A modified Kolmogorov-Smirnov test sensitive to tail alternatives. *The annals of Statistics*, pages 933–946.

Menezes, A. F., Mazucheli, J., and Chakraborty, S. (2021). A collection of parametric modal regression models for bounded data. *Journal of Biopharmaceutical Statistics*, 31(4):490–506.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278.

Meredith, M., Kruschke, J., and Meredith, M. M. (2018). Package 'hdinterval'. *Highest (Posterior) Density Intervals*.

Murray, I., Adams, R., and MacKay, D. (2010). Elliptical slice sampling. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 541–548. JMLR Workshop and Conference Proceedings.

Müller, P. (1991). A generic approach to posterior integration and Gibbs sampling. *Technical report, Purdue University, West Lafayette, Indiana*.

Müller, P. (1993). Alternatives to the Gibbs sampling scheme. *Technical report, Institue of Statistics and Decison Sciences, Duke University*.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77(1).

Novick, S. J. and Stefanski, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458):472–481.

Oakes, D. (1999). Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):479–482.

Ota, H., Kato, K., and Hara, S. (2019). Quantile regression approach to conditional mode estimation. *Electronic Journal of Statistics*, 13(2):3120–3160.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria.

Pratiwi, N., Iswahyudi, C., and Safitri, R. I. (2019). Generalized extreme value distribution for value at risk analysis on gold price. *Journal of Physics: Conference Series*, 1217(1):012090.

Price, C., Wood, M., Leonard, C., Towler, S., Ward, J., Montijo, H., Kellison, I., Bowers, D., Monk, T., Newcomer, J., and et al. (2010). Entorhinal cortex volume in older adults: Reliability and validity considerations for three published measurement protocols. *Journal of the International Neuropsychological Society*, 16:846–855.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239.

Riutort-Mayol, G., Bürkner, P.-C., Andersen, M. R., Solin, A., and Vehtari, A. (2023). Practical hilbert space approximate bayesian gaussian processes for probabilistic programming. *Statistics and Computing*, 33(1):17.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(3):429–453.

Rubio, F. and Steel, M. (2015). Bayesian modelling of skewness and kurtosis with two-piece scale and shape distributions. *Electronic Journal of Statistics*, 9(2):1884–1912.

Sager, T. W. and Thisted, R. A. (1982). Maximum likelihood estimation of isotonic modal regression. *The Annals of Statistics*, pages 690–707.

Schwartz, J. (1999). Air pollution and hospital admissions for heart disease in eight us counties. *Epidemiology*, pages 17–22.

Sheather, S. and Jones, C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(3):683–690.

Shi, J., Zhang, Y., Yu, P., and Song, W. (2021). Simex estimation in parametric modal regression with measurement error. *Computational Statistics & Data Analysis*, 157:107158.

Shin, J. Y., Chen, S., and Kim, T.-W. (2015a). Application of bayesian markov chain monte carlo method with mixed gumbel distribution to estimate extreme magnitude of tsunamigenic earthquake. *KSCE Journal of Civil Engineering*, 19(2):366–375.

Shin, J.-Y., Lee, T., and Ouarda, T. B. M. J. (2015b). Heterogeneous mixture distributions for modeling multisource extreme rainfalls. *Journal of Hydrometeorology*, 16(6):2639–2657.

Siegel, A. F. (2012). *Practical Business Statistics*. Elsevier.

Sitek, G. (2016). The modes of a mixture of two normal distributions. *Silesian Journal of Pure and Applied Mathematics*, 6(1):59–67.

Smith, R. L. (2003). Statistics of extremes, with applications in environment, insurance, and finance. *Extreme Values in Finance, Telecommunications, and the Environment*, pages 20–97.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). BUGS 0.5: Bayesian inference using Gibbs sampling manual (version ii). *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*, pages 1–59.

Stan Development Team (2021). RStan: the R interface to Stan. R package version 2.21.3.

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function a normal mean with application to measurement err oorf models. *Communications in Statistics-Theory and Methods*, 18(12):4335–4358.

Stefanski, L. A., Novick, S. J., and Devanarayan, V. (2005). Estimating a nonlinear function of a normal mean. *Biometrika*, 92(3):732–736.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269.

Tian, M., He, J., and Yu, K. (2017). Fitting truncated mode regression model by simulated annealing. In *Computational Optimization in Engineering-Paradigms and Applications*. Intech Open.

Ullah, A., Wang, T., and Yao, W. (2021). Modal regression for fixed effects panel data. *Empirical Economics*, 60(1):261–308.

Van de Bovenkamp, F. S., Hafkenscheid, L., Rispens, T., and Rombouts, Y. (2016). The emerging importance of igg fab glycosylation in immunity. *The Journal of Immunology*, 196(4):1435–1441.

Varon, D., Barker, W., Loewenstein, D., Greig, M., Bohorquez, A., Santos, I., Shen, Q., Harper, M., Vallejo-Luces, T., and and, R. D. (2014). Visual rating and volumetric measurement of medial temporal atrophy in the alzheimer's disease neuroimaging initiative (ADNI) cohort: baseline diagnosis and the prediction of MCI outcome. *International Journal of Geriatric Psychiatry*, 30(2):192–200.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: an improved r for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718.

Venter, J. (1967). On estimation of the mode. *The Annals of Mathematical Statistics*, 38(5):1446–1455.

Vidal, I. (2014). A bayesian analysis of the gumbel distribution: An application to extreme rainfall data. *Stochastic Environmental Research and Risk Assessment*, 28(3):571–582.

Wang, B. and Shi, J. Q. (2014). Generalized gaussian process regression model for non-gaussian functional data. *Journal of the American Statistical Association*, 109(507):1123–1133.

Wang, H. J., Stefanski, L. A., and Zhu, Z. (2012). Corrected-loss estimation for quantile regression with covariate measurement errors. *Biometrika*, 99(2):405.

Wang, K. and Li, S. (2021). Robust distributed modal regression for massive data. *Computational Statistics & Data Analysis*, 160:107225.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Wei, Y. and Carroll, R. J. (2009). Quantile regression with measurement error. *Journal of the American Statistical Association*, 104(487):1129–1143.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1).

Xiang, S. and Yao, W. (2022a). Modal regression for skewed, truncated, or contaminated data with outliers. In *Advances and Innovations in Statistics and Data Science*, pages 257–273. Springer.

Xiang, S. and Yao, W. (2022b). Nonparametric statistical learning based on modal regression. *Journal of Computational and Applied Mathematics*, 409:114130.

Xu, J., Wang, F., Peng, Q., You, X., Wang, S., Jing, X.-Y., and Chen, C. P. (2020). Modal-regression-based structured low-rank matrix recovery for multiview learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(3):1204–1216.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214.

Yang, H. and Yang, J. (2014). A robust and efficient estimation and variable selection method for partially linear single-index models. *Journal of Multivariate Analysis*, 129:227–242.

Yao, W. and Li, L. (2013). A new regression model: Modal linear regression. *Scandinavian Journal of Statistics*, 41(3):656–671.

Yao, W. and Li, L. (2014). A new regression model: Modal linear regression. *Scandinavian Journal of Statistics*, 41(3):656–671.

Yao, W., Lindsay, B. G., and Li, R. (2012). Local modal regression. *Journal of nonparametric statistics*, 24(3):647–663.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification*. Springer New York.

Yu, K. and Aristodemou, K. (2012). Bayesian mode regression. *arXiv preprint arXiv:1208.0579*.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Yu, K. and Zhang, J. (2005). A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics—Theory and Methods*, 34(9-10):1867–1879.

Zhang, R., Zhao, W., and Liu, J. (2013). Robust estimation and variable selection for semiparametric partially linear varying coefficient model based on modal regression. *Journal of Nonparametric Statistics*, 25(2):523–544.

Zhang, T., Kato, K., and Ruppert, D. (2021). Bootstrap inference for quantile-based modal regression. *Journal of the American Statistical Association*, pages 1–13.

Zhao, W., Zhang, R., Liu, J., and Lv, Y. (2014). Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression. *Annals of the Institute of Statistical Mathematics*, 66(1):165–191.

Zhao, Y., Cheng, Z., Lu, Y., Chang, X., Chan, C., Bai, Y., Zhang, Y., and Cheng, N. (2017). Pm10 and pm2. 5 particles as main air pollutants contributing to rising risks of coronary heart disease: A systematic review. *Environmental Technology Reviews*, 6(1):174–185.

Zhou, H. and Huang, X. (2016). Nonparametric modal regression in the presence of measurement error. *Electronic Journal of Statistics*, 10(2).

Zhou, H. and Huang, X. (2020). Parametric mode regression for bounded responses. *Biometrical Journal*, 62(7):1791–1809.

Zhou, H. and Huang, X. (2022). Bayesian beta regression for bounded responses with unknown supports. *Computational Statistics & Data Analysis*, 167:107345.

Zhou, H., Huang, X., and Initiative, A. D. N. (2020). Parametric mode regression for bounded responses. *Biometrical Journal*, 62(7):1791–1809.

Appendix A

Bayesian modal regression based on mixture distributions

A.1 Proofs of main results

A.1.1 Preliminary lemmas

Before proving the main theorems and propositions, we first prove the following two lemmas.

Lemma A.1.1. Let $p(x)$ be the pdf of an inverse gamma distribution with shape and scale parameters $a \in (0, \infty)$ and $b \in (0, \infty)$ respectively. Then for $p \leq n$,

$$\int_0^\infty \left\{ \frac{\Gamma(0.5x + 0.5)}{\Gamma(0.5x)} \right\}^{n-p} x^{0.5p-0.5n} p(x)\, dx < \infty.$$

Proof. First, let us split the integral into two parts,

$$\begin{aligned} & \int_0^\infty \left\{ \frac{\Gamma(0.5x + 0.5)}{\Gamma(0.5x)} \right\}^{n-p} x^{0.5p-0.5n} p(x)\, dx \\ = & \int_0^1 \left\{ \frac{\Gamma(0.5x + 0.5)}{\Gamma(0.5x)} \right\}^{n-p} x^{0.5p-0.5n} p(x)\, dx \\ & + \int_1^\infty \left\{ \frac{\Gamma(0.5x + 0.5)}{\Gamma(0.5x)} \right\}^{n-p} x^{0.5p-0.5n} p(x)\, dx \\ := & I_1 + I_2. \end{aligned} \tag{A.1.1}$$

We next consider the integrals I_1 and I_2 separately.

Because $\Gamma(x)$ is strictly decreasing for $x \in (0, 1)$, we have

$$\Gamma(0.5x + 0.5) < \Gamma(0.5x), \quad \forall x \in (0, 1).$$

Therefore,

$$\begin{aligned} I_1 &= \int_0^1 \left\{ \frac{\Gamma\left(0.5x+0.5\right)}{\Gamma\left(0.5x\right)} \right\}^{n-p} x^{0.5p-0.5n} p\left(x\right) dx \\ &< \int_0^1 x^{0.5p-0.5n} p\left(x\right) dx \\ &\propto \int_0^1 x^{-(a+0.5n-0.5p)-1} \exp\left(-b/x\right) dx \\ &< \infty, \end{aligned} \tag{A.1.2}$$

where the last line of the display is because $x^{-(a+0.5n-0.5p)-1} \exp\left(-b/x\right)$ is the kernel of an inverse gamma distribution with $a + 0.5n - 0.5p$ and b as the shape and scale parameter respectively.

Finally, we consider I_2. By Gautschi's inequality,

$$\frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s}, \quad \forall x > 0,\, 0 < s < 1,$$

thus

$$\frac{\Gamma(0.5x+0.5)}{\Gamma(0.5x)} < 0.5^{0.5}(x+1)^{0.5}, \quad \forall x > 1.$$

Hence,

$$\begin{aligned} I_2 = &\int_1^\infty \left\{ \frac{\Gamma\left(0.5x+0.5\right)}{\Gamma\left(0.5x\right)} \right\}^{n-p} x^{0.5p-0.5n} p\left(x\right) dx \\ &< \int_1^\infty 0.5^{0.5n-0.5p} \left(x+1\right)^{0.5n-0.5p} x^{0.5p-0.5n} p\left(x\right) dx \\ &\propto \int_1^\infty \left(1+1/x\right)^{0.5n-0.5p} p\left(x\right) dx \\ &< \int_1^\infty 2^{0.5n-0.5p} p\left(x\right) dx \\ &< 2^{0.5n-0.5p} \\ &< \infty. \end{aligned} \tag{A.1.3}$$

Combining (A.1.1)-(A.1.3), one proves the assertion. □

Lemma A.1.2. Let $\boldsymbol{U}_{p\times p} = [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_p]^\top$ be a nonsingular design matrix with finite entries. If $f_Z(y-\theta)$ is the pdf of a distribution from the location family with $\theta \in \mathbb{R}$ as the location parameter, then

$$\int_{\mathbb{R}^p} \prod_{i=1}^p f_Z\left(y_i - \boldsymbol{U}_i^\top \boldsymbol{\beta}\right) d\boldsymbol{\beta} = 1/|\det(\boldsymbol{U})|.$$

Proof. For $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^\top$, let

$$\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{\beta}.$$

Since $\boldsymbol{U}$ is a nonsingular design matrix with finite entries,

$$\boldsymbol{U}^{-1}\boldsymbol{\theta} = \boldsymbol{\beta}.$$

Hence, the corresponding Jacobian matrix of the one-to-one transformation is

$$\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}} = \boldsymbol{U}^{-1}.$$

Using a change of variables with $\theta_i = \boldsymbol{U}_i^\top \boldsymbol{\beta}$, we have that

$$\begin{aligned}
\int_{\mathbb{R}^p} \prod_{i=1}^{p} f_Z\left(y_i - \boldsymbol{U}_i^\top \boldsymbol{\beta}\right) d\boldsymbol{\beta} &= \int_{\mathbb{R}^p} \prod_{i=1}^{p} f_Z\left(y_i - \theta_i\right) \left|\det\left(\boldsymbol{U}^{-1}\right)\right| d\boldsymbol{\theta} \\
&= \left|\det\left(\boldsymbol{U}^{-1}\right)\right| \int_{\mathbb{R}^p} \prod_{i=1}^{p} f_Z\left(y_i - \theta_i\right) d\boldsymbol{\theta} \\
&= \left|\det\left(\boldsymbol{U}^{-1}\right)\right| \\
&= 1/|\det(\boldsymbol{U})|.
\end{aligned}$$

This completes the proof. □

A.1.2 Proofs of Theorems 4.1 and 4.2 and Propositions 4.1 and 4.2

Proof of Theorem 4.4.1

Proof. Recall that $f_Z\left(y \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right)$ has $y = 0$ as the global mode such that

$$f_Z\left(0 \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) \geq f_Z\left(y \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right), \quad \forall y, w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2.$$

Therefore,

$$\begin{aligned}
\prod_{i=1}^{n} f\left(y_i \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) &= \prod_{i=1}^{n} f_Z\left(y_i - \theta \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) \\
&\leq f_Z\left(y_1 - \theta \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) f_Z^{n-1}\left(0 \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right).
\end{aligned}$$

Using this upper bound for the likelihood, we have, for the posterior distribution,

$$\begin{aligned}
&p(w,\theta,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2|\boldsymbol{X},\boldsymbol{Y}) \\
&\propto \left\{\prod_{i=1}^{n} f\left(y_i \mid w,\theta,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right)\right\} p(w)p(\theta)p(\boldsymbol{\xi}_1)p(\boldsymbol{\xi}_2) \\
&\leq f_Z\left(y_1-\theta \mid w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right) f_Z^{n-1}\left(0 \mid w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right) p(w)p(\theta)p(\boldsymbol{\xi}_1)p(\boldsymbol{\xi}_2).
\end{aligned}$$

We next integrate the preceding expression with respect to θ, then with respect to $(w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$ to check the propriety of $p(w,\theta,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2|\boldsymbol{X},\boldsymbol{Y})$.

Taking integration with respect to θ and using the change of variables method, we have that

$$\int_{-\infty}^{+\infty} f_Z\left(y_1-\theta \mid w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right) f_Z^{n-1}\left(0 \mid w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right) d\theta = f_Z^{n-1}\left(0 \mid w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right).$$

Finally, by the sufficient condition given in Theorem 4.4.1, we have

$$\iiint_{\Theta_{w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2}} f_Z^{n-1}\left(0 \mid w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right) p(w)p\left(\boldsymbol{\xi}_1\right)p\left(\boldsymbol{\xi}_2\right)\, dw\, d\boldsymbol{\xi}_1\, d\boldsymbol{\xi}_2 < \infty.$$

It follows that

$$\iiint_{\Theta_{w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2}} \int_{-\infty}^{+\infty} p(w,\theta,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2|\boldsymbol{X},\boldsymbol{Y})\, d\theta\, dw\, d\boldsymbol{\xi}_1\, d\boldsymbol{\xi}_2 < \infty.$$

This shows that the posterior distribution is proper. □

Proof of Theorem 4.4.2

Proof. By assumption, the $n \times p$ design matrix $\boldsymbol{X} := [\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n]^\top$ is full rank. Without loss of generality, we assume that the first p rows of $\boldsymbol{X}$ are linearly independent. Define the submatrix $\boldsymbol{U}_{p\times p}$ consisting of the first p rows of $\boldsymbol{X}$.

Using the fact that the GUD family is a unimodal location family, we have that

$$\begin{aligned}
\prod_{i=1}^{n} f\left(y_i \mid w,\boldsymbol{\beta},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right) &= \prod_{i=1}^{n} f_Z\left(y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta} \mid w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right) \\
&\leq \prod_{i=1}^{p} f_Z\left(y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta} \mid w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right) f_Z^{n-p}\left(0 \mid w,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\right).
\end{aligned}$$

Using this upper bound for the likelihood, we have, for the posterior distribution,

$$\begin{aligned}
&p(w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 | \boldsymbol{X}, \boldsymbol{Y}) \\
\propto &\left\{ \prod_{i=1}^{n} f\left(y_i \mid w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) \right\} p(w) p(\boldsymbol{\beta}) p(\boldsymbol{\xi}_1) p(\boldsymbol{\xi}_2) \\
\leq &\prod_{i=1}^{p} f_Z\left(y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) f_Z^{n-p}\left(0 \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) p(w) p(\boldsymbol{\beta}) p(\boldsymbol{\xi}_1) p(\boldsymbol{\xi}_2).
\end{aligned}$$

We next integrate the preceding expression with respect to $\boldsymbol{\beta}$, then with respect to $(w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ to check the propriety of $p(w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 | \boldsymbol{X}, \boldsymbol{Y})$.

By Lemma A.1.2,

$$\begin{aligned}
&\int_{\mathbb{R}^p} \prod_{i=1}^{p} f_Z\left(y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) f_Z^{n-p}\left(0 \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) d\boldsymbol{\beta} \\
=&f_Z^{n-p}\left(0 \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) / |\det(\boldsymbol{U})|.
\end{aligned}$$

Finally, by the sufficient condition given in Theorem 4.4.2, we have

$$1/|\det(\boldsymbol{U})| \iiint_{\Theta_{w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2}} f_Z^{n-p}\left(0 \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) p(w) p\left(\boldsymbol{\xi}_1\right) p\left(\boldsymbol{\xi}_2\right) dw d\boldsymbol{\xi}_1 d\boldsymbol{\xi}_2 < \infty.$$

Since $\boldsymbol{X}$ has finite entries (and thus, so does $\boldsymbol{U}$), it must be that $\det(\boldsymbol{U})$ is a constant. Further, $\boldsymbol{U}$ is nonsingular, so $\det(\boldsymbol{U}) \neq 0$. It follows that

$$\iiint_{\Theta_{w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2}} \int_{\mathbb{R}^p} p(w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \mid \boldsymbol{X}, \boldsymbol{Y}) \, d\boldsymbol{\beta} \, dw \, d\boldsymbol{\xi}_1 \, d\boldsymbol{\xi}_2 < \infty.$$

This shows that the posterior distribution is proper. □

Proof of Proposition 4.4.1

Proof. The proof consists of verifying the sufficient condition in Theorem 4.4.2 for the three considered Bayesian modal regression models.

First, we show that the modal regression model based on the FG distribution in (4.4.2) has a proper posterior distribution. By Theorem 4.4.2, we need to show that

$$\int_0^\infty \int_0^\infty \int_0^1 f_{\mathrm{FG}}^{n-p}\left(y = 0 \mid w, \theta = 0, \sigma_1, \sigma_2\right) p(w) p(\sigma_1) p(\sigma_2) dw d\sigma_1 d\sigma_2 < \infty.$$

Note that

$$\begin{aligned} f_{\text{FG}}\left(y=0 \mid w, \theta=0, \sigma_1, \sigma_2\right) &= \exp(-1)\left(w/\sigma_1+(1-w)/\sigma_2\right) \\ &\leq \exp(-1)\left(1/\sigma_1+1/\sigma_2\right). \end{aligned}$$

Therefore, it is sufficient to show

$$\begin{aligned} &\int_0^\infty \int_0^\infty \int_0^1 \left(1/\sigma_1+1/\sigma_2\right)^{n-p} p(w) p\left(\sigma_1\right) p\left(\sigma_2\right) dw d\sigma_1 d\sigma_2 \\ =&\int_0^\infty \int_0^\infty \left(1/\sigma_1+1/\sigma_2\right)^{n-p} p\left(\sigma_1\right) p\left(\sigma_2\right) d\sigma_1 d\sigma_2 \\ =&\int_0^\infty \int_0^\infty \sum_{k=0}^{n-p} \binom{n-p}{k} \left(1/\sigma_1\right)^{n-p-k} \left(1/\sigma_2\right)^k p\left(\sigma_1\right) p\left(\sigma_2\right) d\sigma_1 d\sigma_2 \\ <&\infty. \end{aligned}$$

The last inequality is true because we use the inverse gamma distribution as the prior for σ_1 and σ_2 and because for any inverse gamma random variable X, $\mathbb{E}[1/X^k] < \infty$ for all $k \in \mathbb{N}$.

Second, we want to show that the linear modal regression model based on the DTP-Student-t distribution (4.4.3) has a proper posterior distribution. We have

$$\begin{aligned} f_{\text{DTP-Student-}t}\left(y=0 \mid \theta=0, \sigma_1, \sigma_2, \delta_1, \delta_2\right) &= 2\left(1-w\right) \frac{\Gamma\left(0.5\delta_2+0.5\right)}{\Gamma(0.5\delta_2)} \frac{1}{\sqrt{\delta_2 \pi} \sigma_2} \\ &\leq 2 \frac{\Gamma\left(0.5\delta_2+0.5\right)}{\Gamma(0.5\delta_2)} \frac{1}{\sqrt{\delta_2 \pi} \sigma_2} \\ &< 2 \frac{\Gamma\left(0.5\delta_2+0.5\right)}{\Gamma(0.5\delta_2)} \frac{1}{\sqrt{\delta_2} \sigma_2}, \end{aligned}$$

where $w \in [0,1]$ is defined in (4.3.6). Applying Lemma A.1.1 and the fact that for any inverse gamma random variable X, $\mathbb{E}[1/X^k] < \infty$ for all $k \in \mathbb{N}$, we have

$$\begin{aligned} &\int_0^\infty \int_0^\infty \left\{ \frac{\Gamma\left(0.5\delta_2+0.5\right)}{\Gamma\left(0.5\delta_2\right)} \frac{1}{\sqrt{\delta_2} \sigma_2} \right\}^{n-p} p(\sigma_2) p(\delta_2) d\sigma_2 d\delta_2 \\ =& \int_0^\infty \left[\int_0^\infty \left\{ \frac{\Gamma\left(0.5\delta_2+0.5\right)}{\Gamma\left(0.5\delta_2\right)} \right\}^{n-p} \delta_2^{0.5p-0.5n} p(\delta_2) d\delta_2 \right] \left(\frac{1}{\sigma_2}\right)^{n-p} p(\sigma_2) d\sigma_2 \\ <& \infty. \end{aligned}$$

Therefore, by Theorem 4.4.2, the posterior distribution for regression model in (4.4.3) is proper.

Lastly, one can show that the linear modal regression model based on the TPSC-Student-t distribution (4.4.4) also has a proper posterior distribution. The proof is almost identical to the proof of posterior propriety for the DTP-Student-t distribution and is therefore omitted. □

Proof of Proposition 4.4.2

Proof. Recall that the pdf of the GUD family is

$$f\left(y \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) = w f_1\left(y \mid \theta, \boldsymbol{\xi}_1\right) + (1-w) f_2\left(y \mid \theta, \boldsymbol{\xi}_2\right).$$

Without loss of generality, we assume that $\tau \in \boldsymbol{\xi}_1$ and $\tau \notin \boldsymbol{\xi}_2$. Suppose that there is only one observation, we have that

$$\int_{\tau \in \Theta_\tau} p\left(w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \mid y\right) d\tau \geq \int_{\tau \in \Theta_\tau} (1-w) f_2\left(y \mid \theta, \boldsymbol{\xi}_2\right) p\left(\tau\right) d\tau = \infty,$$

since $p(\tau)$ is improper.

When there is more than one observation, binomial expansion of the GUD likelihood gives $\prod_{i=1}^{n} f\left(y_i \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) \geq C[f_2\left(y \mid \theta, \boldsymbol{\xi}_2\right)]^n$, where $C > 0$ is free of $\boldsymbol{\xi}_1$. Hence, for $C[f_2\left(y \mid \theta, \boldsymbol{\xi}_2\right)]^n p\left(\tau\right)$, the integration with respect to τ is still divergent when $p(\tau)$ is improper. □

A.2 The lognormal mixture distribution

To demonstrate how researchers can add new members to the GUD family, we present the construction of the lognormal mixture distribution (logNM) below. Here, we pick the lognormal distribution because the lognormal distribution is right-skewed and unimodal. We construct a location-shift lognormal distribution such that the transformed lognormal distribution is still right-skewed but has a mode at 0. Next, we flip the location-shift lognormal distribution at 0 to get a left-skewed unimodal lognormal distribution. Finally, we mix the left- and right-skewed lognormal distribu-

tion together to construct the logNM distribution. More details of the construction of logNM can be found in (A.2.1)-(A.2.2).

The pdf of the lognormal distribution is

$$f_{\text{logN}}(y \mid \mu, \nu) = \frac{1}{y\nu\sqrt{2\pi}} \exp\left(-\frac{(\ln(y)-\mu)^2}{2\nu^2}\right) \mathbb{I}(y > 0), \tag{A.2.1}$$

where $\mu \in (-\infty, +\infty)$ and $\nu > 0$ are two parametrers, and the mode is given by $\exp(\mu - \nu^2)$. We define the pdf of logNM as a mixture of two lognormal pdfs formulated as follows,

$$\begin{aligned} f_{\text{logNM}}(y \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) &= w f_1(y|\theta, \boldsymbol{\xi}_1) + (1-w) f_2(y|\theta, \boldsymbol{\xi}_2), \\ f_1(y|\theta, \boldsymbol{\xi}_1) &= f_{\text{logN}}\left(\exp\left(\mu_1 - \nu_1^2\right) - (y-\theta) \mid \mu_1, \nu_1\right), \\ f_2(y|\theta, \boldsymbol{\xi}_2) &= f_{\text{logN}}\left(\exp\left(\mu_2 - \nu_2^2\right) + (y-\theta) \mid \mu_2, \nu_2\right), \end{aligned} \tag{A.2.2}$$

where $\boldsymbol{\xi}_1 = [\mu_1, \nu_1]^\top$ and $\boldsymbol{\xi}_2 = [\mu_2, \nu_2]^\top$. It can be shown that both component distributions in (A.2.2) are unimodal at θ, with continuous densities over the real line, one left-skewed and the other right-skewed. Moreover, the pdfs of the individual lognormal mixture components in (A.2.2) have 0 at the right and left boundaries of their supports. Hence, the pdf of the logNM distribution is continuous in all of $\mathbb{R}$. Having verified that all three restrictions (R1)-(R3) for the GUD family (introduced in Section 3 of the main manuscript), we can proceed to use the logNM likelihood (A.2.3) for Bayesian modal regression.

Figure A.1 demonstrates that the logNM distributions can be asymmetric or symmetric given different combinations of parameter values. The top panel shows that, with an increase of μ_2, the right tail of the logNM distribution becomes heavier while its left tail remains almost the same. The bottom panel shows how ν_1 influences the amount of skewness of the logNM distribution when all three logNM distributions are left-skewed.

Practitioners can build a Bayesian modal linear regression model based on the

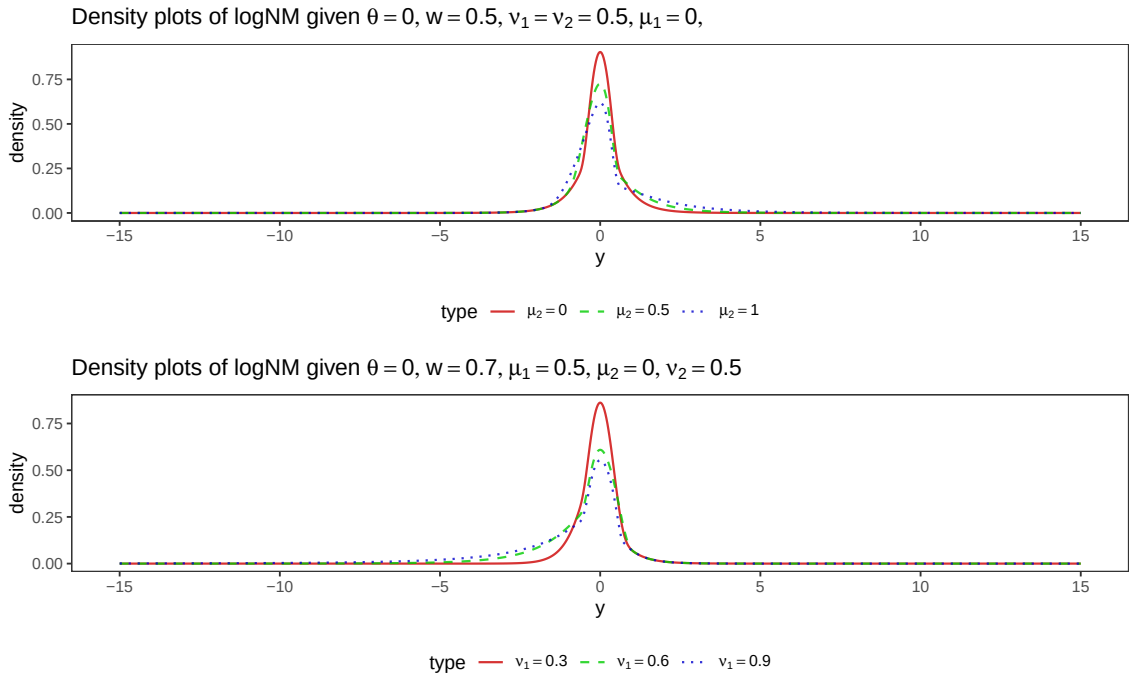


Figure A.1 Density plots of the logNM distribution given different combinations of parameter values.

logNM likelihood (A.2.3) as follows:

$$
\begin{aligned}
Y_i \mid \boldsymbol{X}_i, w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 &\overset{\text{ind}}{\sim} \text{logNM}\left(w, \boldsymbol{X}_i^\top \boldsymbol{\beta}, \mu_1, \nu_1, \mu_2, \nu_2\right), \\
w &\sim \text{Uniform}(0,1), \\
\nu_1, \nu_2 &\overset{\text{i.i.d}}{\sim} \text{InverseGamma}(1,1), \\
\mu_1, \mu_2 &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 100^2), \\
p(\boldsymbol{\beta}) &\propto \mathcal{N}_p\left(\mathbf{0}, 10^2 \times \boldsymbol{I}_{p\times p}\right),
\end{aligned}
\tag{A.2.3}
$$

where $\boldsymbol{I}_{p\times p}$ stands for the p by p identity matrix and $\boldsymbol{\beta}$ is a p-dimension random vector. If one wishes to use a flat prior $p(\boldsymbol{\beta}) \propto 1$, then one must verify the sufficient condition in Theorem 4.4.2 of chapter 4.

If researchers have another right-skewed or a left-skewed distribution to work with, then they can mimic the construction of the logNM above to propose a member of the GUD family that works for their applications. For example, one can use the reparameterized unimodal right-skewed Gamma distribution from Bourguignon et al.

(2020) to construct a new type I GUD and a corresponding modal regression model.

In the following lemma, we show why flat priors *cannot* be used for μ_1 and(or) μ_2 in the Bayesian modal regression model based on the logNM distribution. This simple example demonstrates why we should *avoid* placing improper priors on the *non*-location parameters in Bayesian modal regression models based on the GUD family.

Proposition A.2.1. Endowing μ_1 and(or) μ_2 with flat priors $p(\mu_1) \propto 1$ and(or) $p(\mu_2) \propto 1$ leads to an improper posterior distribution under the logNM model (A.2.3).

Proof. We want to show that

$$\int_{-\infty}^{+\infty} \prod_{i=1}^{n} f_{\text{logNM}}\left(y_i \mid w, \theta, \sigma_1, \sigma_2, \mu_1, \mu_2\right) d\mu_1 = +\infty,$$

and(or),

$$\int_{-\infty}^{+\infty} \prod_{i=1}^{n} f_{\text{logNM}}\left(y_i \mid w, \theta, \sigma_1, \sigma_2, \mu_1, \mu_2\right) d\mu_2 = +\infty.$$

Since

$$f_{\text{logNM}}\left(y \mid w, \theta, \sigma_1, \sigma_2, \mu_1, \mu_2\right) = w f_{\text{logN}}\left(\exp\left(\mu_1 - \nu_1^2\right) - (y - \theta) \mid \theta, \mu_1, \nu_1\right) + (1 - w) f_{\text{logN}}\left(\exp\left(\mu_2 - \nu_2^2\right) + (y - \theta) \mid \theta, \mu_2, \nu_2\right),$$

and any pdf must be nonnegative, it suffices to show that

$$\int_{-\infty}^{+\infty} f_{\text{logN}}\left(\exp\left(\mu_1 - \nu_1^2\right) - (y - \theta) \mid \theta, \mu_1, \nu_1\right) d\mu_2 = \infty,$$

and(or)

$$\int_{-\infty}^{+\infty} f_{\text{logN}}\left(\exp\left(\mu_2 - \nu_2^2\right) + (y - \theta) \mid \theta, \mu_2, \nu_2\right) d\mu_1 = \infty.$$

Both the integrals above are non-finite. This completes the proof. □

Following the same arguments as those in Proposition A.2.1, one can show that improper priors such as $p(\nu_1) \propto 1/\nu_1$ and(or) $p(\nu_2) \propto 1/\nu_2$ will also lead to an improper posterior distribution. As stated in Proposition 4.2 in chapter 4, a general rule is that, for the Bayesian modal regression models based on the GUD family,

using improper prior(s) for any parameter in $(\boldsymbol{\xi}_1 \cup \boldsymbol{\xi}_2) \setminus (\boldsymbol{\xi}_1 \cap \boldsymbol{\xi}_2)$ leads to an improper posterior distribution. Here, $A \backslash B = A \cap B^c$ denotes a collection of elements in A but not in B.

A.3 A short note about Markov Chain Monte Carlo (MCMC)

Readers who are familiar with Bayesian modeling of mixture distributions may wonder why we do not use the data augmentation "trick" to design a specific MCMC algorithm for modal regression models based on the GUD family. The problem lies in the type II distributions of the GUD family. If $\mathcal{D}_1 \cap \mathcal{D}_2 = \varnothing$, then the latent variable conditional on other parameters and observed data becomes a degenerate random variable. This degenerate random variable will not behave randomly.

To demonstrate that we have a degenerate random variable, let us consider a simple case with a single observation. Recall that the type II GUD has the pdf

$$f\left(y \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) = w f_1\left(y \mid \theta, \boldsymbol{\xi}_1\right) I\left(y < \theta\right) + (1-w) f_2\left(y \mid \theta, \boldsymbol{\xi}_2\right) I\left(y \geq \theta\right).$$

Introducing the latent variable z, we have the joint pdf as

$$\begin{aligned} f\left(y, z \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) =& \left[w f_1\left(y \mid \theta, \boldsymbol{\xi}_1\right) I(y < \theta)\right]^z \\ & \left[(1-w) f_2\left(y \mid \theta, \boldsymbol{\xi}_2\right) I(y \geq \theta)\right]^{1-z}. \end{aligned}$$

The conditional distribution of z is then

$$\begin{aligned} p\left(z \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, y\right) &\propto f\left(y, z \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) \\ &\sim \text{Bernoulli}\left(r\right), \end{aligned}$$

where

$$r = \frac{w f_1\left(y \mid \theta, \boldsymbol{\xi}_1\right) I\left(y < \theta\right)}{w f_1\left(y \mid \theta, \boldsymbol{\xi}_1\right) I\left(y < \theta\right) + (1-w) f_2\left(y \mid \theta, \boldsymbol{\xi}_2\right) I\left(y \geq \theta\right)}.$$

Similarly, the conditional distribution of θ is

$$p\left(\theta \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, y, z\right) \propto f\left(y, z \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) p\left(\theta\right).$$

The conditional mean of the latent variable z can only be 0 *or* 1 since $y < \theta$ and $y \geq \theta$ cannot happen at the same time. Hence, the latent variable becomes a degenerate random variable during the MCMC iterations. Without loss of generality, let us assume $z = 1$ in the first iteration of the MCMC. It is not hard to see that, during the MCMC iterations, the updated values of θ can only be larger than y. This leads to z being equal to 1 for the rest of the iterations in the MCMC algorithm. However, if the true value of θ is smaller than y, then the MCMC chain will never reach the true value. Similarly, if $z = 0$ in the first iteration of the MCMC, then all updated values of θ can only be smaller than y.

In conclusion, for Bayesian modal regression models based on the type II GUD subfamily, the latent variable data augmentation algorithm will *not* explore the *whole* parameter space. Hence, it is *non*-ergodic. As a consequence of this, we do *not* use the data augmentation "trick" for mixture models in our MCMC algorithm. Instead, we use the No-U-Turn Sampler implemented in the Stan software (Hoffman et al., 2014; Carpenter et al., 2017).

Convergence Diagnostics for the Real Data Applications and Simulation Studies

In this section, we include more details about posterior inference, convergence diagnostics, and traceplots for the four data application examples and two simulation studies from the main manuscript. The `rhat`, which has the theoretical minimum value as 1, is a statistic measuring the convergence of the MCMC chains. To obtain reliable posterior inference, it is recommended that `rhat` should be near 1 or at least less than 1.1 (page 287 of Gelman et al. (2013)). The `ess_bulk` and `ess_tail` are the bulk and tail effective sample size respectively. The `ess_tail` is defined as the minimum of the effective sample sizes for the 5% and 95% quantiles. The recommended lower threshold for `ess_bulk` and `ess_tail` is 400 (Vehtari et al., 2021). All of the `rhat`,

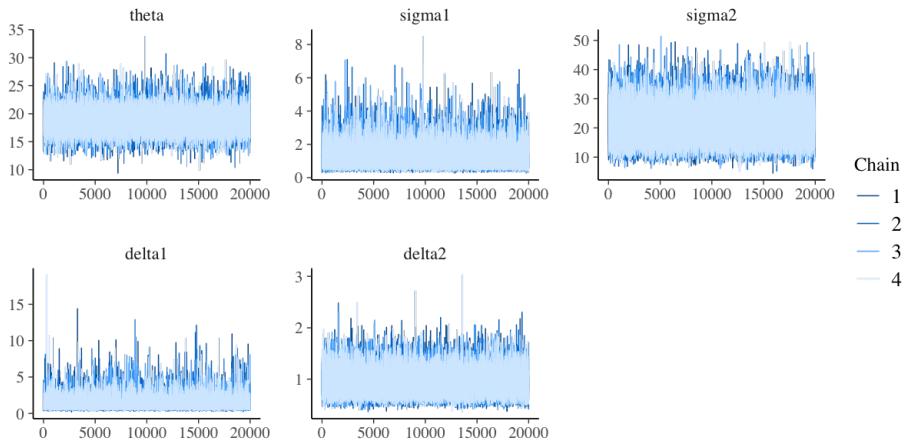


Figure A.2 Traceplots for the modal regression model based on the DTP-Student-t likelihood fit to the bank deposits data.

`ess_bulk` and `ess_tail` summary statistics from our analyses in the main manuscript satisfy these recommended thresholds.

All of the traceplots presented in this section also indicate that the MCMC chains mixed well, with no convergence issues. We ran four MCMC chains for each model that was fit and used the combined MCMC samples to approximate the posterior distributions. For the intercept-only regression model fit to the bank deposits data (Section 2.1 of the main manuscript), we set the number of warmup iterations as 10,000 and the number of post-warmup iterations as 20,000 for each chain. For all other models in chapter 4, we set the number of warmup iterations as 10,000 and the number of post-warmup iterations as 10,000 for each of the four MCMC chains.

Bank Deposits Application from Section 4.2.1

```
# A tibble: 6 × 10
variable     mean   median    sd   mad       q5     q95  rhat ess_bulk ess_tail
<chr>       <dbl>    <dbl> <dbl> <dbl>    <dbl>   <dbl> <dbl>    <dbl>    <dbl>
1 lp__     -288.    -288.    1.63  1.43  -291.    -286.    1.00   36112.   47263.
2 sigma1      1.20     1.08  0.559 0.437    0.567    2.25  1.00   51221.   45698.
3 sigma2     20.5     20.0   5.37  5.20    12.6     30.1   1.00   46745.   49890.
4 delta1      1.34     1.16  0.745 0.503    0.597    2.67  1.00   70792.   47613.
5 delta2      0.915    0.888 0.211 0.198    0.621    1.30  1.00   51342.   54038.
6 theta      18.7     18.7   1.80  1.67    15.9     21.6   1.00   42904.   40976.
```

Crime Rate Application from Section 4.2.2

```
## mean regression - Normal likelihood

# A tibble: 4 × 10
variable     mean   median     sd    mad       q5     q95  rhat ess_bulk ess_tail
<chr>       <dbl>    <dbl>  <dbl>  <dbl>    <dbl>   <dbl> <dbl>    <dbl>    <dbl>
1 alpha    -24.2    -24.2    5.33   5.28   -33.0    -15.4    1.00   15362.   20037.
2 beta[1]    0.467    0.466  0.163  0.161    0.200    0.738  1.00   17845.   20507.
3 beta[2]    1.14     1.14   0.227  0.225    0.765    1.51   1.00   18553.   22035.
4 beta[3]    0.0677   0.0677 0.0341 0.0340   0.0121   0.124  1.00   23524.   24312.

## median regression - ALD likelihood

# A tibble: 4 × 10
variable    mean  median     sd    mad       q5    q95  rhat ess_bulk ess_tail
<chr>      <dbl>   <dbl>  <dbl>  <dbl>   <dbl>  <dbl> <dbl>    <dbl>    <dbl>
1 alpha    -1.34   -1.18   3.39   3.28   -7.16   3.89    1.00    9895.   11638.
2 beta[1]  -0.118  -0.123  0.0964 0.0933 -0.268  0.0471  1.00   10933.   13580.
3 beta[2]   0.437   0.432  0.138  0.128   0.216  0.673   1.00   11652.   13166.
4 beta[3]   0.0555  0.0554 0.0161 0.0158  0.0294 0.0818  1.00   17520.   18274.

## modal regression - TPSC-Student-t

# A tibble: 4 × 10
variable    mean  median     sd    mad       q5     q95  rhat ess_bulk ess_tail
<chr>      <dbl>   <dbl>  <dbl>  <dbl>    <dbl>   <dbl> <dbl>    <dbl>    <dbl>
1 alpha     1.12    1.25   2.68   2.62   -3.56     5.29    1.00   13814.   16690.
2 beta[1]  -0.199  -0.204  0.0825 0.0805 -0.327   -0.0552  1.00   14342.   17771.
3 beta[2]   0.243   0.250  0.138  0.144   0.00688  0.459   1.00   10233.   15650.
4 beta[3]   0.0636  0.0624 0.0153 0.0149  0.0403   0.0906  1.00   10145.    9351.
```

Air Pollution Application from Section 4.6

```
## mean regression - Normal likelihood

# A tibble: 2 × 10
variable  mean median    sd   mad    q5    q95  rhat ess_bulk ess_tail
<chr>    <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>    <dbl>    <dbl>
1 alpha    41.9   41.9  3.12  3.14  36.8  47.1    1.00   15822.   19673.
2 beta[1]  -1.27  -1.26 0.841 0.849 -2.64  0.116  1.00   15614.   18993.


## median regression - ALD likelihood

# A tibble: 2 × 10
variable  mean median    sd   mad    q5   q95  rhat ess_bulk ess_tail
<chr>    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
1 alpha    32.8   32.7  2.05  2.03  29.5  36.2   1.00   13984.   16088.
2 beta[1]  -1.87  -1.85 0.530 0.524 -2.77 -1.03  1.00   14043.   15873.

## modal regression - TPSC-Student-t
```

```
# A tibble: 2 × 10
variable  mean median    sd   mad    q5    q95  rhat ess_bulk ess_tail
<chr>    <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>    <dbl>    <dbl>
1 alpha     9.67  9.76  1.44  1.18   7.46 11.7    1.00    4268.    1787.
2 beta[1]  -1.01 -0.997 0.309 0.302 -1.53 -0.538  1.00    4581.    1832.
```

Serum Data Application from Section 4.6.2

```
## mean regression - Normal likelihood

# A tibble: 3 × 10
variable    mean  median     sd    mad     q5    q95  rhat ess_bulk ess_tail
<chr>      <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <dbl>    <dbl>    <dbl>
1 alpha     3.08    3.08   0.384  0.380   2.45  3.72    1.00   10706.   15240.
2 beta[1]   0.969   0.969  0.313  0.311   0.451 1.48    1.00    9820.   13370.
3 beta[2]  -0.0458 -0.0460 0.0510 0.0508 -0.130 0.0385  1.00   10328.   14088.


## median regression - ALD likelihood

# A tibble: 3 × 10
variable    mean  median     sd    mad     q5    q95  rhat ess_bulk ess_tail
<chr>      <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <dbl>    <dbl>    <dbl>
1 alpha     2.82    2.81   0.441  0.449   2.12  3.57    1.00    8105.   12034.
2 beta[1]   1.12    1.13   0.348  0.352   0.527 1.68    1.00    7692.   10288.
3 beta[2]  -0.0656 -0.0668 0.0574 0.0576 -0.157 0.0314  1.00    8134.   10823.

## modal regression - TPSC-Student-t

# A tibble: 3 × 10
variable   mean median     sd    mad     q5     q95  rhat ess_bulk ess_tail
<chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl> <dbl>    <dbl>    <dbl>
1 alpha     2.37   2.37  0.320  0.316   1.84   2.89    1.00   12565.   17546.
2 beta[1]   1.15   1.16  0.264  0.264   0.720  1.59    1.00   11055.   16023.
3 beta[2]  -0.107 -0.107 0.0441 0.0438 -0.179 -0.0338  1.00   11692.   17288.
```

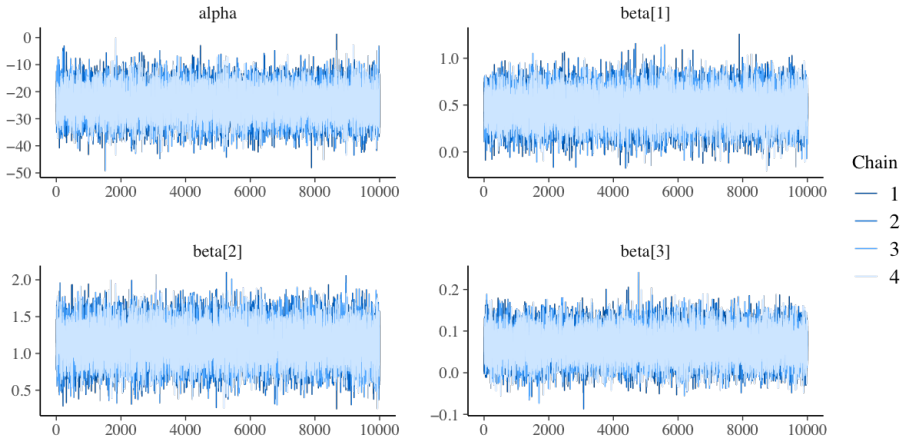
Left-Skewed Simulation Study from Section 4.5.1

```
## mean regression - Normal likelihood

# A tibble: 2 × 10
variable  mean median    sd   mad    q5   q95  rhat ess_bulk ess_tail
<chr>    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
1 alpha    -2.32  -2.32  2.37  2.33 -6.21  1.59  1.00   36206.   27772.
2 beta[1]   2.64   2.66  3.67  3.59 -3.36  8.66  1.00   34513.   27175.


## median regression - ALD likelihood

# A tibble: 2 × 10
variable  mean median    sd   mad      q5   q95  rhat ess_bulk ess_tail
<chr>    <dbl>  <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>    <dbl>    <dbl>
```

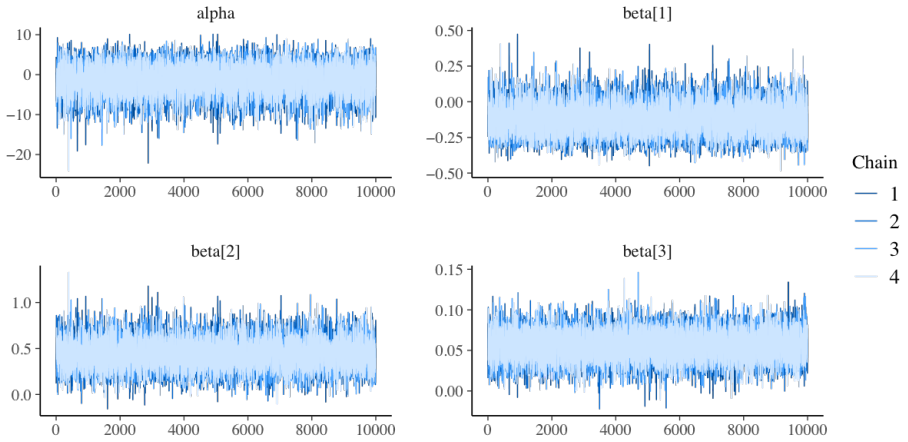
```
1 alpha    0.653  0.648 0.468 0.445 -0.0964  1.43  1.00   31916.   25876.
2 beta[1]  0.751  0.746 0.755 0.693 -0.483   2.00  1.00   34392.   25226.

## modal regression - TPSC-Student-t

# A tibble: 2 × 10
variable  mean median    sd   mad     q5   q95  rhat ess_bulk ess_tail
<chr>    <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>    <dbl>
1 alpha    0.847  0.819 0.386 0.364 0.265   1.52  1.00   20058.   19143.
2 beta[1]  0.639  0.644 0.334 0.321 0.0800  1.18  1.00   29384.   25974.
```

Right-Skewed Simulation Study from Section 4.5.2

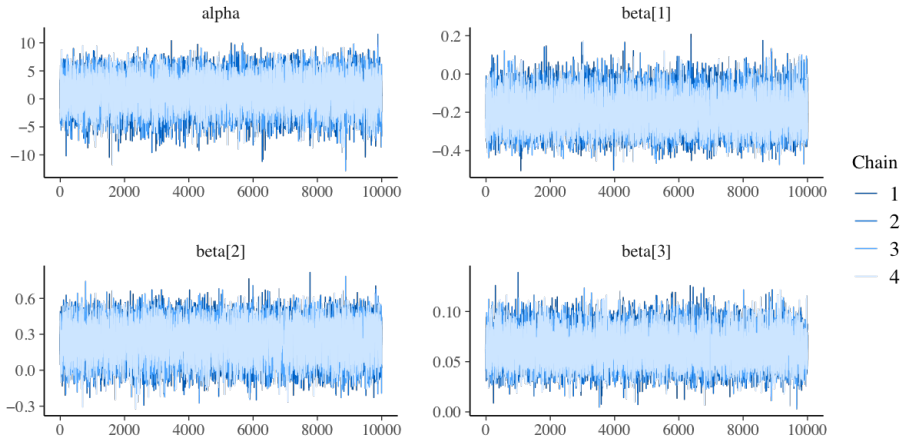
```
## mean regression - Normal likelihood

# A tibble: 2 × 10
variable  mean median    sd   mad    q5   q95  rhat ess_bulk ess_tail
<chr>    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>
1 alpha     1.83   1.82  2.01  1.97 -1.46  5.12  1.00   37079.   27813.
2 beta[1]   3.12   3.11  3.43  3.35 -2.51  8.76  1.00   34832.   26126.


## median regression - ALD likelihood

# A tibble: 2 × 10
variable  mean median    sd   mad     q5   q95  rhat ess_bulk ess_tail
<chr>    <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>    <dbl>
1 alpha    0.976  0.986 0.470 0.464  0.192  1.73  1.00   30065.   23935.
2 beta[1]  0.473  0.487 0.944 0.925 -1.12   1.98  1.00   28958.   26648.

## modal regression - TPSC-Student-t

# A tibble: 2 × 10
variable  mean median    sd   mad     q5   q95  rhat ess_bulk ess_tail
<chr>    <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>    <dbl>
1 alpha    1.33   1.41  0.532 0.517  0.328  2.05  1.00   14412.   15244.
2 beta[1]  0.257  0.211 0.589 0.612 -0.633  1.28  1.00   20263.   24407.
```



(a) Traceplots for the mean regression model

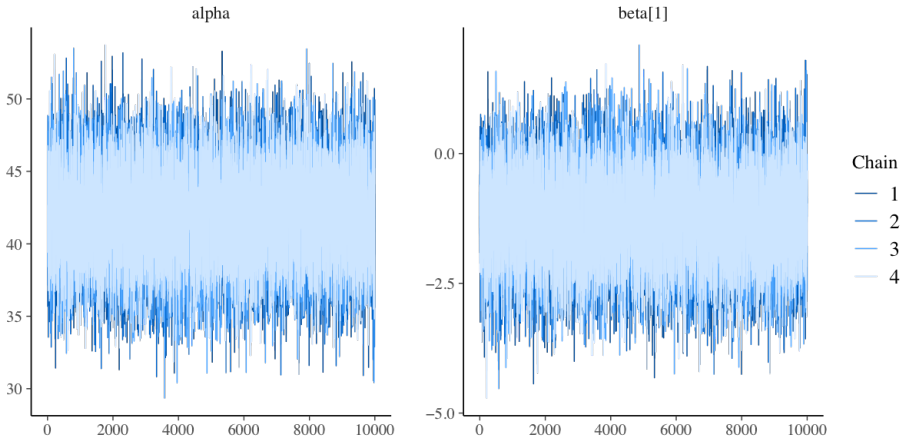


(b) Traceplots for the median regression model

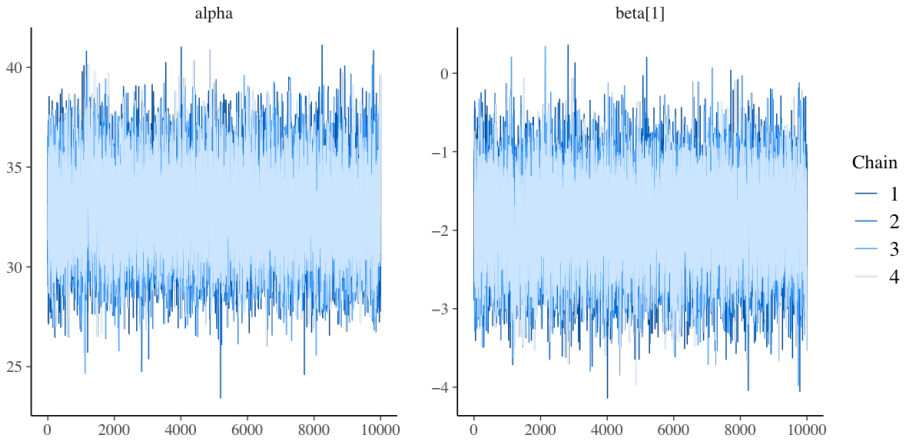


(c) Traceplots for the modal regression model based on the TPSC-Student-t likelihood

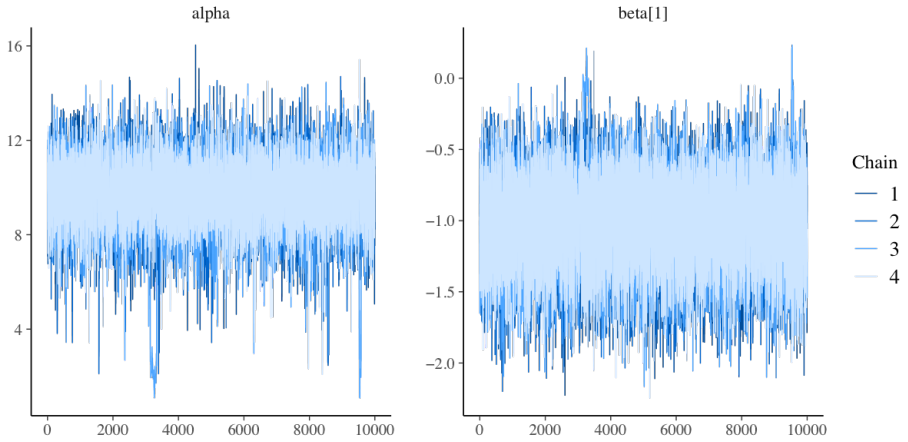
Figure A.3 Traceplots for the mean/median/modal regression models fit to the crime data.



(a) Traceplots for the mean regression model

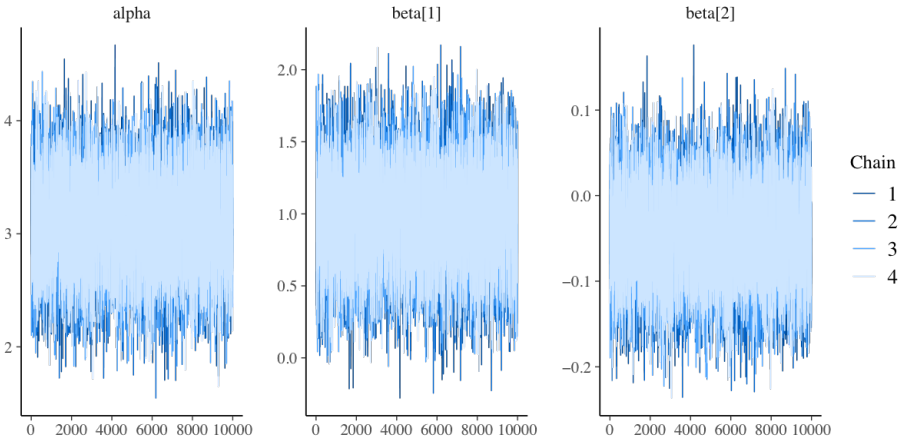


(b) Traceplots for the median regression model

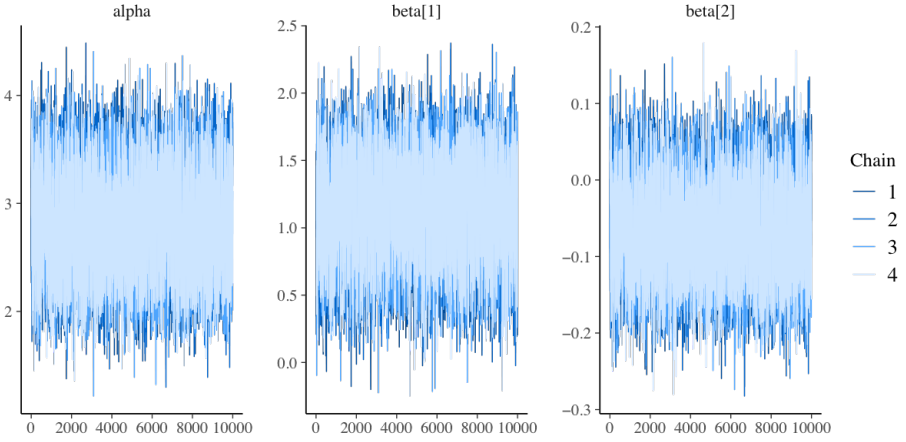


(c) Traceplots for the modal regression model based on the TPSC-Student-t likelihood

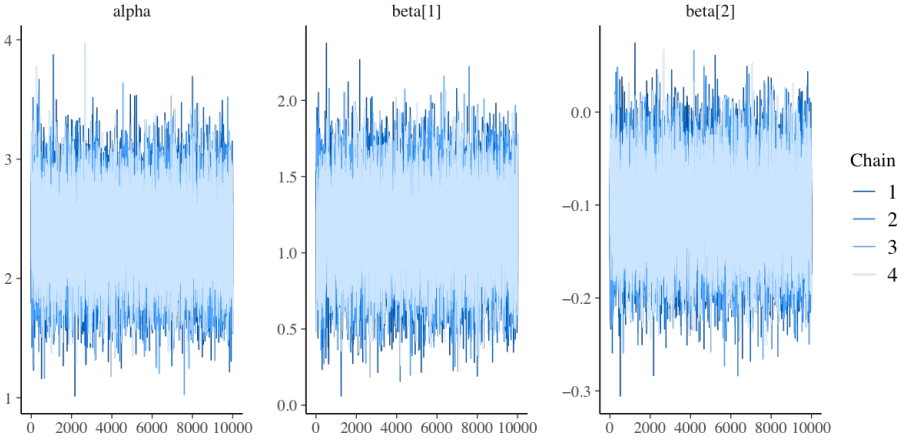
Figure A.4 Traceplots for the mean/median/modal regression models fit to the air pollution data.



(a) Traceplots for the mean regression model

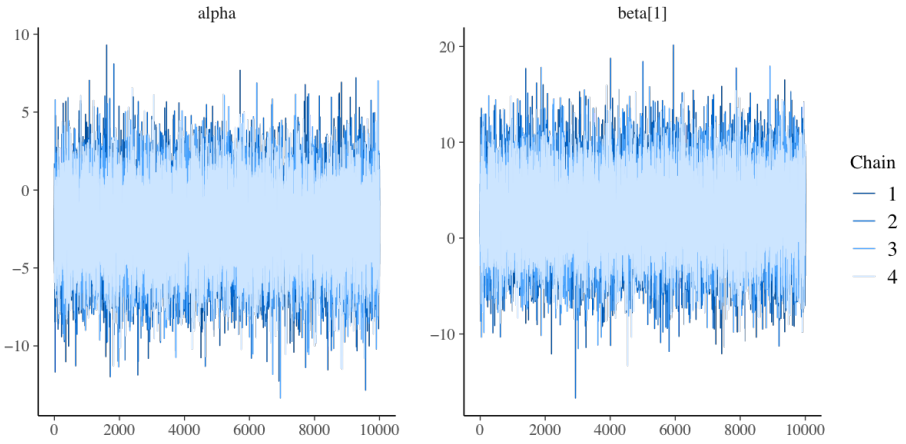


(b) Traceplots for the median regression model

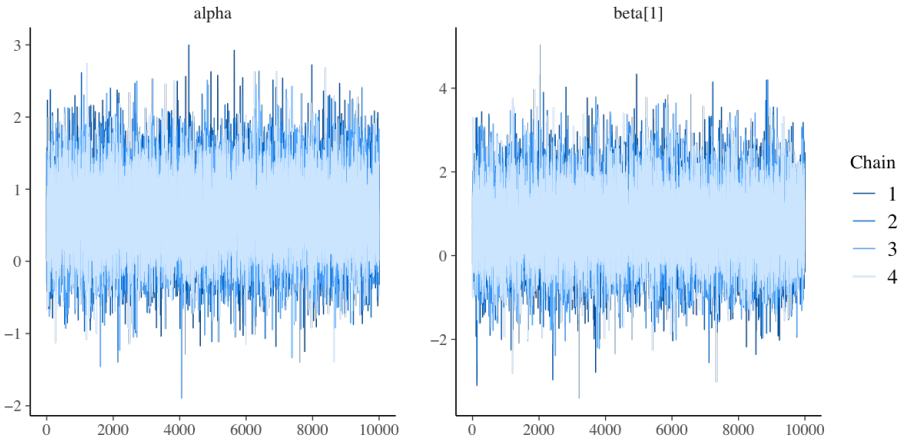


(c) Traceplots for the modal regression model based on the TPSC-Student-t likelihood

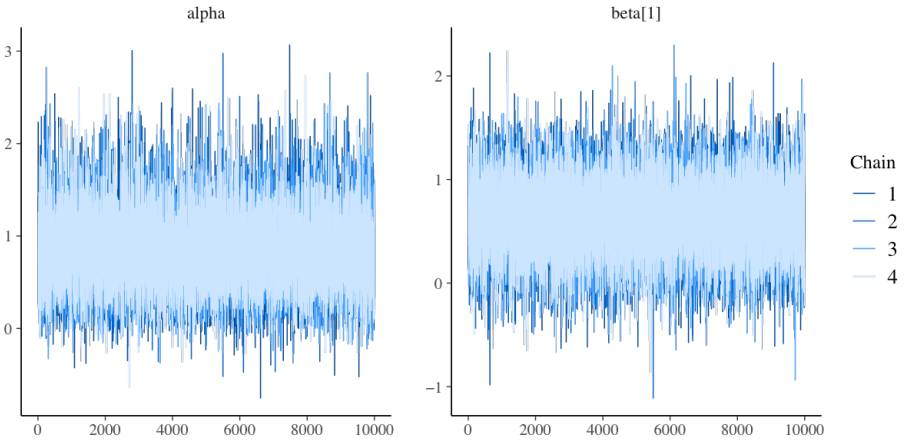
Figure A.5 Traceplots for the mean/median/modal regression models fit to the serum data.



(a) Traceplots for the mean regression model

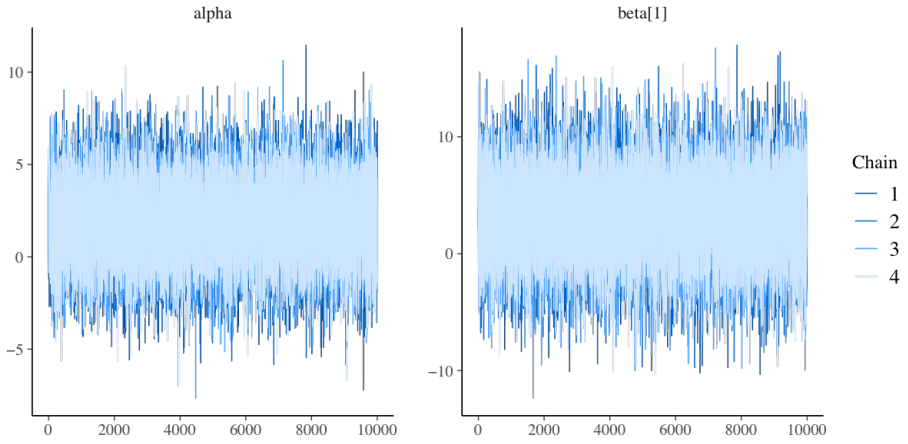


(b) Traceplots for the median regression model

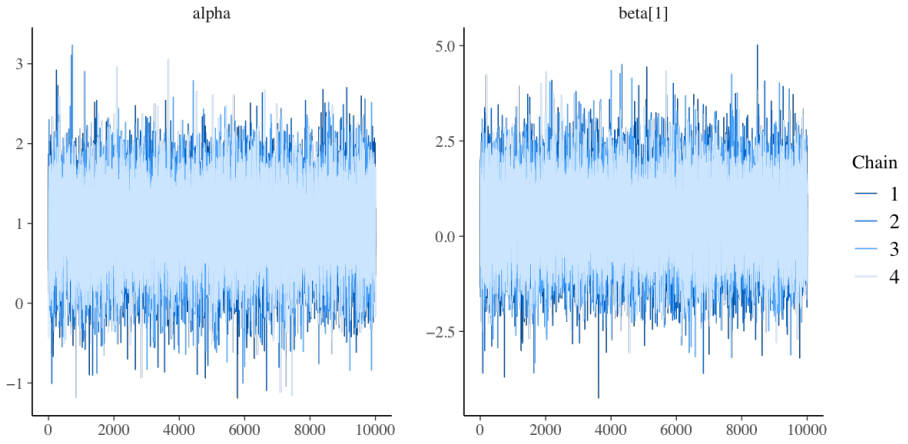


(c) Traceplots for the modal regression model based on the TPSC-Student-t likelihood

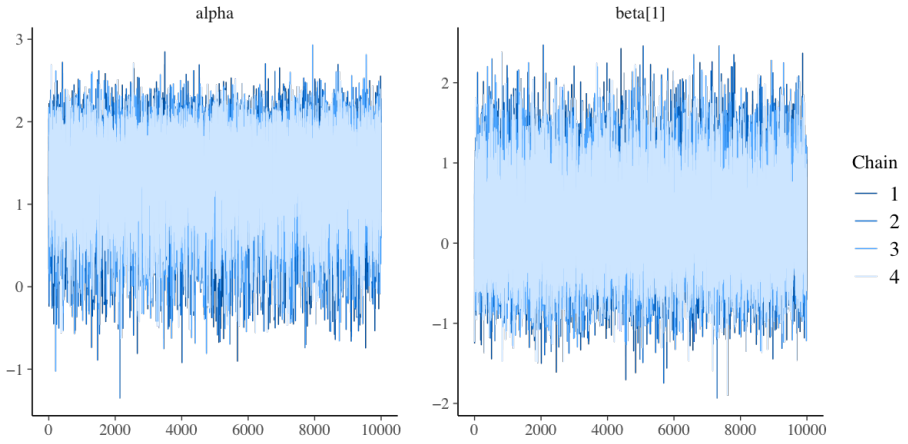
Figure A.6 Traceplots for the mean/median/modal regression models from the left-skewed simulation study.



(a) Trraceplots for the mean regression model



(b) Traceplots for the median regression model



(c) Traceplots for the modal regression model based on the TPSC-Student-t likelihood

Figure A.7 Traceplots for the mean/median/modal regression models from the right-skewed simulation study.