Bayesian Modal Regression Based on Mixture Distributions

Qingyang Liu[a], Xianzheng Huang[a], Ray Bai[a]

[a]*Department of Statistics, University of South Carolina, Columbia, 29201, South Carolina, United States*

Abstract

Compared to mean regression and quantile regression, the literature on modal regression is very sparse. A unifying framework for Bayesian modal regression is proposed that is based on a family of unimodal distributions indexed by the mode along with other parameters that allow for flexible shapes and tail behaviors. Sufficient conditions for posterior propriety under an improper prior on the mode parameter are derived. Following prior elicitation, regression analysis of simulated data and datasets from several real-life applications are conducted. Besides drawing inference for covariate effects that are easy to interpret, prediction and model selection under the proposed Bayesian modal regression framework are also considered. Evidence from these analyses suggest that the proposed inference procedures are very robust to outliers, enabling one to discover interesting covariate effects missed by mean or median regression, and to construct much tighter prediction intervals than those from mean or median regression. Computer programs for implementing the proposed Bayesian modal regression are available at https://github.com/rh8liuqy/Bayesian_modal_regression.

Keywords:
Mode, Fat-tailed distribution, Outlier, Unimodal distribution, Robust statistics

1. Introduction

There is an abundance of literature on mean regression models which model the conditional mean of a response variable Y given a set of covariates $\boldsymbol{X}$. However, it is no secret that the mean is sensitive to outliers. Median regression – or more generally, quantile regression – is robust to outliers and

is thus an appealing alternative to mean regression (Koenker et al., 2017). Besides the mean and median, the mode is yet another commonly used measure of central tendency. Compared with mean or median regression, modal regression concerns the conditional *mode* of Y given $\boldsymbol{X}$ and is much less explored (Sager and Thisted, 1982; Lee, 1989, 1993), especially in the parametric framework.

But why are modal regression models 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, 2022). 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. 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. By construction, modal regression explores the relationship between the "most probable" value of Y and $\boldsymbol{X}$, and thus offers a highly interpretable representative value of the response. For example, in precipitation forecasting, it is easier and of greater public interest to apprehend the most likely predicted rainfall (i.e. the mode of precipitation amount) rather than the predicted mean or median rainfall (Dalenius, 1965).

A major challenge in building parametric modal regression models is constructing an appropriate distribution family that subsumes asymmetric, symmetric, light-tailed, and fat-tailed mode-zero error distributions to flexibly model the distribution of Y given $\boldsymbol{X}$, with the location parameter being the $\boldsymbol{X}$-dependent mode. In this paper, we propose the general unimodal distribution (GUD) family, which is a subfamily of the general two-component mixture distribution family described in Section 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 3.4. Such flexibility can provide a higher level of protection from model misspecification than mean regression, which requires at least the first moment of the error distribution to exist. 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.

1.1. Existing work on modal regression

Frequentist nonparametric modal regression has been the mainstream in the limited existing literature on modal regression (Yao and Li, 2014; Chen et al., 2016; Ota et al., 2019). For readers interested in frequentist nonparametric modal regression, we refer to 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 and Huang (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 positive or 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 Aristodemou (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. Ho 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.

1.2. Our contributions

This paper 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. The main contributions of this paper 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 selecting an appropriate member of the GUD family 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, real estate, and molecular biology: a) robustness to outliers, b) precise prediction, and c) high interpratability of covariate effects.

By accomplishing the first three items above, we provide a comprehensive toolkit for modal regression analysis based on a flexible family containing a large variety of unimodal distributions. Our fully Bayesian approach allows researchers to easily infer covariate effects through their posterior distributions instead of relying on asymptotic approximations or resampling methods like the bootstrap (Boos and Stefanski, 2013). In a similar spirit, constructing prediction intervals based on the posterior predictive distribution also becomes straightforward, especially with a simple algorithm for sampling from the GUD family (discussed in Section 3).

The structure of this paper is as follows. In Section 2, we motivate our proposed Bayesian modal regression framework with two data applications. In Section 3, we formally define the GUD family and zoom in on several important members in the family. Section 4 introduces Bayesian inference for these modal regression models, including prior elicitation, posterior propriety,

uncertainty quantification, and model selection. Section 5 provides simulation studies that illustrate the strengths of our methodology. In Section 6, we provide two additional data applications from real estate and molecular biology. Section 7 concludes the paper with some remarks about our Bayesian modal regression framework and several directions for future research.

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 3 and 4) to datasets from the economics and criminology literature.

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 (2016)).

Figure 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 3) to the dataset, along with the histogram of the observed 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 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 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

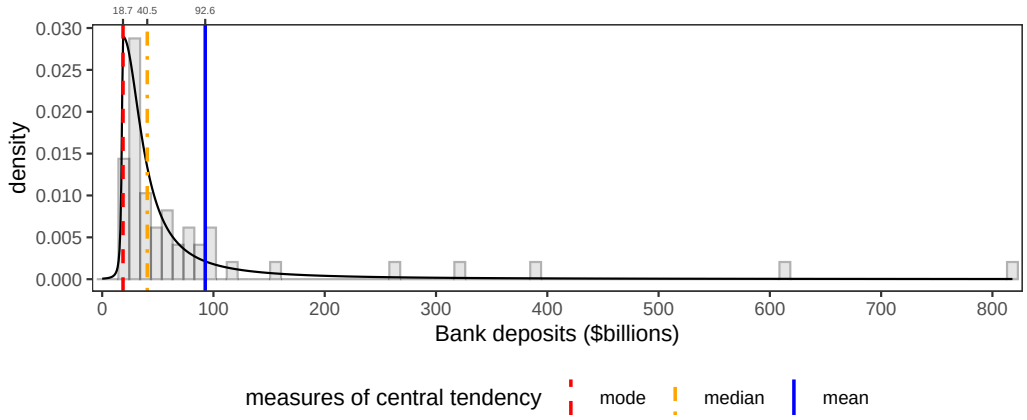

Figure 1: *Deposits (in billions of dollars) of 50 banks and savings institutions in the United States on July 2, 2010. The solid black curve is the estimated density of the DTP-Student-t distribution. The three vertical lines mark locations of the sample mean (blue solid line), the sample median (orange dash-dotted line), and the estimated mode (red dashed line), respectively.*

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 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 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.

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 first created the conditional scatter plot matrix of the complete data (see the left panel in Figure 2), in which D.C. stands out as an outlier. This outlier is extreme enough to

visually obscure potential association between the considered variables. We thus created a second conditional scatter plot matrix after removing D.C. (see the right panel in Figure 2). Now one can more easily perceive a positive association between the poverty percentage and the murder rate, as well as a positive association between the metropolitan rate and the murder rate. Although a visual inspection seems to suggest a negative association between the college percentage and the murder rate, a more thorough regression analysis of the murder rates data is needed to confirm this.

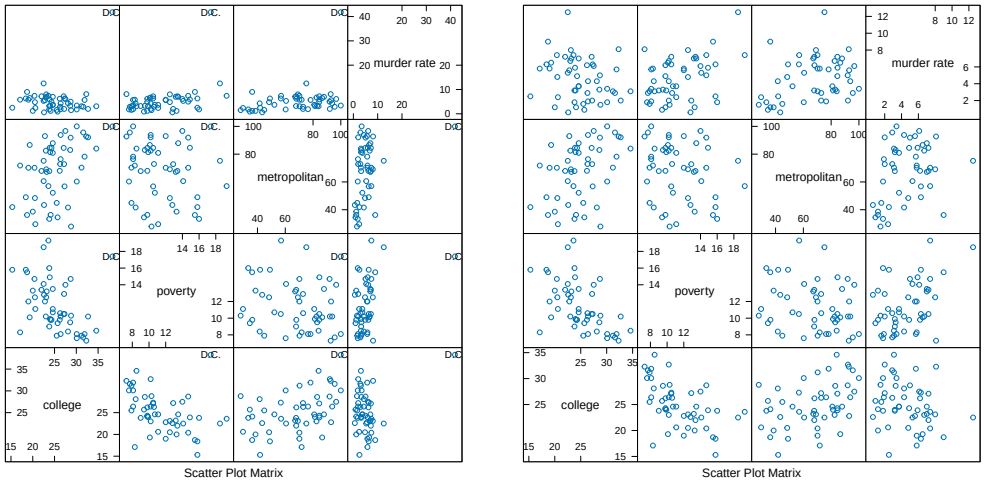

Figure 2: *The conditional scatter plot matrices of U.S. crime data, with the left matrix including D.C. that is labeled in the plot, and the right matrix excluding D.C.*

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 1 presents the inference results from mean/median/modal regression models with D.C. included in the data (see the upper half of Table 1) and the counterpart results with D.C. excluded (see the lower half of Table 1). Some of the results are shared by all three models in both rounds of analyses. Namely, all models determine that the poverty percentage and the metropolitan rate

were both positively associated with the murder rate. However, the first round of analysis (using the complete data) results in different conclusions about the association between the college percentage and the murder rate. With a 90% posterior credible interval (CI) of $(0.20, 0.74)$, the mean regression model (specified by (14)) implies that there exists a *positive* association between the college percentage and the murder 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 90% CI of $(-0.27, 0.05)$, the Bayesian median regression model (formulated in (16)) 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 3) draws a different conclusion. With a 90% 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 4.3) in Table 1, the modal regression model based on the TPSC-Student-t likelihood yields the highest value of ELPD, indicating a better fit to the data than the mean and median regression models.

With the D.C. outlier removed from the data in the second round of analysis, the median and modal regression models do in fact suggest a significant negative association between the college percentage and the murder rate, with CIs of $(-0.29, -0.06)$ and $(-0.32, -0.06)$, respectively. Even though the point estimate for the effect of the college percentage on the murder rate in the mean regression model is now negative, the CI of this covariate effect is $(-0.27, 0.02)$, and thus one would not conclude it as a significant effect according to mean regression. In fact, Table 1 shows that under the mean regression model, both point estimates and interval estimates for most of the covariate effects change substantially in the second round of analysis. In contrast, the results from our second round of modal regression analysis mostly remain the same as those from the first round. This exercise demonstrates that modal regression based on the proposed GUD family can be even more robust to outliers than median regression and has the potential to draw reliable inferences and capture important features of data even in the presence of extreme outliers.

Table 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.

D.C. inclusion	Regression Model	ELPD	Parameter(covariate)	Mean	q5	q95
D.C. included	Mean regression	-161.81	β_1(college)	0.47	0.20	0.74
			β_2(poverty)	1.14	0.77	1.52
			β_3(metropolitan)	0.07	0.01	0.12
	Median regression	-133.42	β_1(college)	-0.12	-0.27	0.05
			β_2(poverty)	0.44	0.21	0.68
			β_3(metropolitan)	0.06	0.03	0.08
	Modal regression	-123.24	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
D.C. excluded	Mean regression	-109.54	β_1(college)	-0.13	-0.27	0.02
			β_2(poverty)	0.35	0.16	0.55
			β_3(metropolitan)	0.06	0.04	0.09
	Median regression	-109.66	β_1(college)	-0.18	-0.29	-0.06
			β_2(poverty)	0.35	0.18	0.52
			β_3(metropolitan)	0.05	0.03	0.08
	Modal regression	-111.94	β_1(college)	-0.20	-0.32	-0.06
			β_2(poverty)	0.24	0.01	0.45
			β_3(metropolitan)	0.06	0.04	0.09

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 2, we now formally introduce the GUD family for Bayesian modal regression.

3.1. The formulation as a mixture distribution

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

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

In the mixture pdf (1), $w \in [0,1]$ is the weight parameter, $\theta \in (-\infty, +\infty)$ is the mode as a location parameter in (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 $f_2(\cdot \mid \theta, \boldsymbol{\xi}_2)$. The supports of the two mixture components, denoted respectively by $\mathcal{D}_1$ and $\mathcal{D}_2$, can be the same or different, as exemplified later in this section. Moreover, the two components are chosen to achieve unimodality and other desirable properties that we elaborate in more detail next. 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 (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 (1) is continuous in its domain.

Restriction (R1) is already implied earlier when we stress that the two components in (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 when samples from f_1 and f_2 are easy to obtain:

(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 sampling 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 4.3.

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 Sections 3.2-3.4, we

detail several examples of distributions from the GUD family. All mixture components in these examples belong to some location-scale family, but f_1 and f_2 are not limited to location-scale families in general (see an example in Section B in the supplementary material). Apart from the GUD family, another class of distributions that can accommodate both skewed/symmetric responses and fat-tailed/thin-tailed responses is the skew-normal family that is well explored by Azzalini (2013). There are two well-received parameterizations of the skew-normal family: the direct parameterization (SNDP) and the centered parameterization (SNCP) (Arellano-Valle and Azzalini, 2008; Durante, 2019). Except for the symmetric members in the SNDP family, the location parameter in an SNDP typically cannot be interpreted as the mean, median, or mode. Meanwhile, the SNCP family is indexed by three parameters, with one of them being the mean. In contrast to SNDP and SNCP, the location parameter of the GUD family is the *mode*, which makes it convenient for drawing inference for the (conditional) mode.

3.2. 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., 2015). 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. (2024) proposed the flexible Gumbel (FG) distribution specified by the pdf

$$f_{\text{FG}}\left(y \mid w, \theta, \sigma_1, \sigma_2\right) = w f_{\text{Gumbel}}\left(-y \mid -\theta, \sigma_1\right) + (1-w) f_{\text{Gumbel}}\left(y \mid \theta, \sigma_2\right). \quad (2)$$

By mapping to (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 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 or weekly heaviest/lightest traffic on a highway.

3.3. 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 $(1/\sigma)g((y-\theta)/\sigma \mid \delta)$ 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

$$f_{\mathrm{LT}}(y \mid \theta, \sigma, \delta) = \frac{2}{\sigma} g\left(\frac{y-\theta}{\sigma}\middle| \delta\right) \mathbb{I}(y < \theta), \tag{3}$$

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

$$f_{\mathrm{RT}}(y \mid \theta, \sigma, \delta) = \frac{2}{\sigma} g\left(\frac{y-\theta}{\sigma}\middle| \delta\right) \mathbb{I}(y \geq \theta). \tag{4}$$

By mixing the pdfs in (3)-(4), we obtain the DTP pdf as

$$f_{\mathrm{DTP}}(y \mid \theta, \sigma_1, \sigma_2, \delta_1, \delta_2) = w f_{\mathrm{LT}}(y \mid \theta, \sigma_1, \delta_1) + (1-w) f_{\mathrm{RT}}(y \mid \theta, \sigma_2, \delta_2), \tag{5}$$

where

$$w = \frac{\sigma_1 g\left(0 \mid \delta_2\right)}{\sigma_1 g\left(0 \mid \delta_2\right) + \sigma_2 g\left(0 \mid \delta_1\right)}, \tag{6}$$

and $g\left(0 \mid \delta\right)$ represents $g((y-\theta)/\sigma \mid \delta)$ evaluated at $y = \theta$. The weight (6) is chosen to guarantee a mixture distribution that satisfies (R3) even though this particular choice makes $w \neq 0, 1$ if $g(0 \mid \delta)$ never vanishes. Restrictions (R1) and (R2) are trivially satisfied by the construction of the left/right θ-truncated pdfs in (3)–(4). Thus, DTP distributions belong to the GUD family. Note, however, that our general GUD family (1) does not *require* the two component densities to be truncated, as we demonstrated earlier with the FG distribution specified by the density in (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 (3) and (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 (5) and (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.

3.4. The two-piece scale distribution

By setting $\delta_1 = \delta_2 = \delta$ in (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,

$$\begin{aligned} f_{\text{TPSC}}\left(y\middle|\, w, \theta, \sigma, \delta\right) =& w f_{\text{LT}}\left(y\middle|\, \theta, \sigma\sqrt{\frac{w}{1-w}}, \delta\right) + \\ & (1-w) f_{\text{RT}}\left(y\middle|\, \theta, \sigma\sqrt{\frac{1-w}{w}}, \delta\right). \end{aligned} \tag{7}$$

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 (7) because we find it more straightforward to elicit a noninformative prior for w than placing a noninformative prior on γ. Interestingly, it is not difficult to show that w follows a uniform distribution on the interval $(0, 1)$ if and only if γ follows a log-logistic distribution with the scale parameter as 1 and the shape parameter as 2 (Ekawati et al., 2015).

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 4.3. All of these models can be easily implemented using the code developed for this work.

3.5. Pictorial depiction of GUD and 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 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 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 $\sigma_1 = \sigma_2$. When the weight parameter w surpasses 0.5, the pdf of FG distribution puts more weight on the left-skewed part, and therefore, becomes more left-skewed. In the third panel, we see that as the scale parameter of the left-skewed component 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. Recall that $\mathcal{D}_1$ 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 (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 Section B of the supplementary material, 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. Lastly, all distributions in GUD considered in this section are constructed by mixing two component distributions belonging to a common distribution family or similar families (e.g. the Gumbel family for constructing the FG distribution). This is done in order to more easily satisfy restriction (R3) of the pdf being continuous

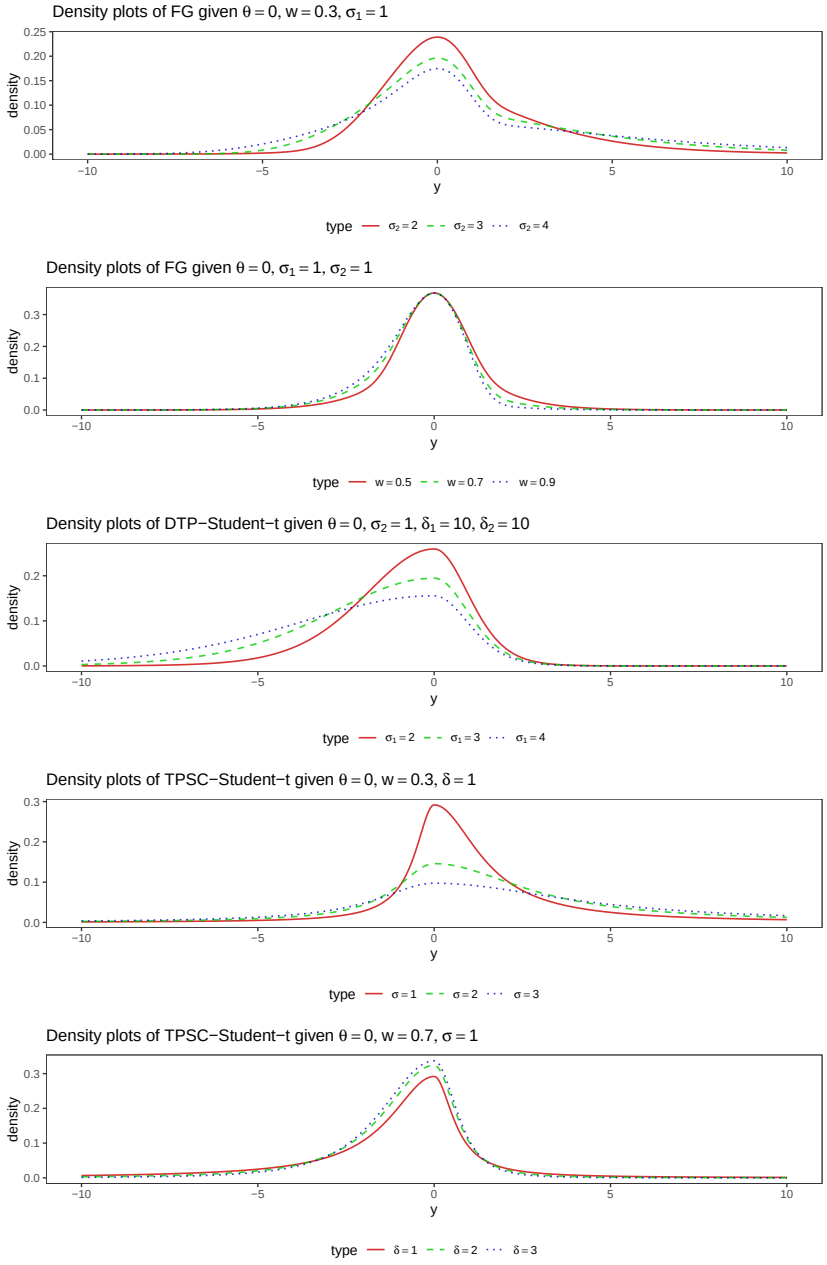

Figure 3: *Density plots of different distributions in the GUD family with different parameter specifications.*

everywhere. More generally, one could start with two components from different families. However, mixing two components from different families can easily lead to a mixture density that is discontinuous at the mode if not carefully constructed.

4. Bayesian modal regression

Having defined the GUD family in Section 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 $\boldsymbol{X}$.

4.1. Prior elicitation

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

$$Y_i \mid \boldsymbol{X}_i, w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \overset{\text{ind}}{\sim} \text{GUD}\left(w,\, \boldsymbol{X}_i^\top \boldsymbol{\beta},\, \boldsymbol{\xi}_1,\, \boldsymbol{\xi}_2\right), \text{ for } i = 1, \ldots, n, \tag{8}$$

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 (8).

To conduct inference for our model in (8), 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 inverse gamma 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 $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ do not share any common parameter, 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 4.2, 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 3, we have the Bayesian modal linear regression model based on the FG likelihood (2) formulated as follows,

$$\begin{aligned} Y_i \mid \boldsymbol{X}_i, w, \boldsymbol{\beta}, \sigma_1, \sigma_2 &\overset{\text{ind}}{\sim} \text{FG}\left(w,\, \boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma_1,\, \sigma_2\right), \text{ for } i = 1, \ldots, n, \\ w &\sim \text{Uniform}(0, 1), \\ \sigma_1, \sigma_2 &\overset{\text{i.i.d}}{\sim} \text{InverseGamma}(1, 1), \\ p(\boldsymbol{\beta}) &\propto 1, \end{aligned} \tag{9}$$

where "i.i.d" refers to "independent and identically distributed." Meanwhile, the Bayesian modal linear regression associated with the DTP-Student-t likelihood (5) is specified by

$$\begin{aligned} Y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta}, \sigma_1, \sigma_2, \delta_1, \delta_2 &\overset{\text{ind}}{\sim} \text{DTP-Student-}t\left(\boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma_1,\, \sigma_2,\, \delta_1,\, \delta_2\right), \text{ for } i = 1, \ldots, n, \\ \sigma_1, \sigma_2, \delta_1, \delta_2 &\overset{\text{i.i.d}}{\sim} \text{InverseGamma}(1, 1), \\ p(\boldsymbol{\beta}) &\propto 1. \end{aligned} \tag{10}$$

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 (7) is defined as

$$\begin{aligned} Y_i \mid \boldsymbol{X}_i, w, \boldsymbol{\beta}, \sigma, \delta &\overset{\text{ind}}{\sim} \text{TPSC-Student-}t\left(w,\, \boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma,\, \delta\right), \text{ for } i = 1, \ldots, n, \\ w &\sim \text{Uniform}(0, 1), \\ \sigma, \delta &\overset{\text{i.i.d}}{\sim} \text{InverseGamma}(1, 1), \\ p(\boldsymbol{\beta}) &\propto 1. \end{aligned} \tag{11}$$

According to Proposition 2 in next subsection, all of the proposed Bayesian modal regression models (9)–(11) above have proper posterior distributions. Practitioners can construct various other Bayesian modal regression models using the same strategy shown above. In this paper, we concentrate on the modal regression models based on the FG, DTP-Student-t, and TPSC-Student-t likelihoods for the sake of concreteness.

4.2. 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 1 gives sufficient conditions under which the GUD likelihood (1) with a flat prior on the mode/location parameter and suitably chosen priors on other model parameters lead to a proper posterior. Theorem 2 extends this result to the regression setting. Proofs for the theorems and propositions in this section can be found in Section A of the supplementary material. 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 (1) as $f_Z(y - \theta \mid w, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = f(y \mid w, \theta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$.

Theorem 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*

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

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 1 applies to the case where there is a single location parameter θ (as in the bank deposits application in Section 2.1). Next, we extend this result to the regression setting. Theorem 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 (1).

Theorem 2. *Let $\boldsymbol{X}$ be a full rank design matrix with $p \leq n$ and finite entries. If*

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

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 Theorems 1 and 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 1. *Suppose that $\boldsymbol{X}$ is full rank with $p \leq n$ and finite entries. Then the Bayesian modal regression models (9), (10), and (11) based on the FG, DTP-Student-t, and TPSC-Student-t likelihoods, respectively, have proper posterior distributions.*

Proposition 1 confirms that under suitable regularity conditions on the design matrix $\boldsymbol{X}$, all of the regression models proposed in this paper have proper posterior distributions. The proof of Proposition 1 relies on verifying the sufficient condition given in Theorem 2. Our proof provides a template for verifying posterior propriety for other Bayesian modal regression models (8) 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 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 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 Section B of the supplementary material, we provide a specific example of Proposition 2 for the logNM distribution (also introduced in the same section).

4.3. Uncertainty quantification and model selection

Let $\boldsymbol{\Omega} = (w, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ represent the collection of model parameters. We are interested in the distribution of a new response Y_{new} given new covariates $\boldsymbol{X}_{\text{new}}$, and observed data $(\boldsymbol{Y}, \boldsymbol{X})$.

The posterior predictive distribution is defined as

$$p\left(Y_{\text{new}} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}}\right) = \int_{\Theta} p\left(Y_{\text{new}} \mid \boldsymbol{\Omega}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}}\right) p\left(\boldsymbol{\Omega} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}}\right) d\boldsymbol{\Omega},$$

where Θ denotes the parameter space. Because the unobserved data is conditionally independent of the observed data $(\boldsymbol{Y}, \boldsymbol{X})$ given the parameters $\boldsymbol{\Omega}$, we have $p(Y_{\text{new}} \mid \boldsymbol{\Omega}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}}) = p(Y_{\text{new}} \mid \boldsymbol{\Omega}, \boldsymbol{X}_{\text{new}})$. Additionally, because the new set of covariates $\boldsymbol{X}_{\text{new}}$ is independent of the posterior distribution of the parameters $\boldsymbol{\Omega}$, $p(\boldsymbol{\Omega} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}}) = p(\boldsymbol{\Omega} \mid \boldsymbol{Y}, \boldsymbol{X})$. Therefore, the posterior predictive distribution reduces to

$$p\left(Y_{\text{new}} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}}\right) = \int_{\Theta} p\left(Y_{\text{new}} \mid \boldsymbol{\Omega}, \boldsymbol{X}_{\text{new}}\right) p\left(\boldsymbol{\Omega} \mid \boldsymbol{Y}, \boldsymbol{X}\right) d\boldsymbol{\Omega}. \tag{12}$$

Obtaining an approximation of the posterior predictive distribution specified by (12) is computationally inexpensive. With the sampling method outlined in Section 3 for the GUD family, one can easily draw samples from $p(Y_{\text{new}} \mid \boldsymbol{\Omega}, \boldsymbol{X}_{\text{new}})$ during each iteration in our MCMC algorithm, and then obtain samples from the posterior predictive distribution $p\left(Y_{\text{new}} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}_{\text{new}}\right)$. In this paper, 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 probability density (HPD) intervals. We use 90% HPD prediction intervals as the posterior prediction intervals for all mean/median/modal regression models that we consider in Sections 5 and 6.

Due to the inherent nature of the conditional mode, the HPD 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). 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 5 and 6.

As mentioned in Section 3, 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

$$\text{ELPD} = \sum_{i=1}^{n} \log p\left(Y_i \mid Y_{-i}\right), \tag{13}$$

where Y_{-i} represents all observations except the i-th observation. In (13), "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 the R package `loo`, which is compatible with the `Stan` programming language (Carpenter et al., 2017). 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 paper.

Other model selection criteria, such as the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) and the widely applicable information criterion (WAIC) introduced by Watanabe and Opper (2010), are also applicable to regression models with GUD likelihoods. In fact, DIC and WAIC have been shown to be asymptotically equal to ELPD (Gelman et al., 2013). However, Vehtari et al. (2017) recommended ELPD and WAIC over DIC because DIC is not a fully Bayesian information criterion and is based on a point estimate. Additionally, Vehtari et al. (2017) demonstrated that ELPD is more robust than WAIC in finite samples with weak priors or influential observations. Therefore, we decided to use ELPD for all data applications and numerical studies in this paper.

4.4. Implementation

We utilized the `Stan` programming language interfaced with R (Carpenter et al., 2017) to implement all data analyses in the empirical study presented in this article. `Stan` uses Hamiltonian Monte Carlo (Neal, 2011) and leverages the No-U-Turn sampler (NUTS) proposed by Hoffman and Gelman (2014). The implementation of Bayesian linear modal regression in (8) in `Stan` involves defining data log-likelihood functions first, followed by specifying priors as outlined in (9) through (11). Computer programs for reproducing all numerical results in our study are available at the following link: https://github.com/rh8liuqy/Bayesian_modal_regression.

Even though it may seem plausible to apply the sampling method as outlined in Section 3 where one introduces a latent variable to separate the inference for the left-skewed and right-skewed parts, dealing with type II GUD distributions introduces challenges. Introducing such a latent variable in a data augmentation MCMC algorithm might result in a degenerate random variable, as demonstrated in Section C of the supplementary material. In such cases, the data augmented MCMC sampler is *reducible*. A Markov chain is reducible if it is impossible to eventually get from one state to any other states in a finite number of steps. In our context, the data augmentation algorithm is reducible since a poor initialization leads to the latent variable always taking the same value, regardless of the number of MCMC iterations. A reducibile MCMC sampler is less practical, since it requires "acceptable" initial conditions in order to be able to explore the entire parameter space (Robert and Casella, 2004). Acceptable initial conditions are typically unknown.

In contrast to many widely used data augmentation algorithms, NUTS does not require the introduction of discrete latent variables to fit Bayesian mixture models. Instead, NUTS simply requires one to be able to calculate the gradient of the log-likelihood with respect to the model parameters, and this can be done using automatic differentiation. We observed satisfactory performance of NUTS in the empirical study, as evidenced by promising convergence results for all data applications and simulations studies (see the convergence diagnostics presented in Section D) of the supplementary material.

5. Simulation studies

5.1. Left-skewed data

We now present one simulation study 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 left-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 classic/robust Bayesian mean and median regression models. The classic Bayesian mean regression used a normal likelihood, i.e.,

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \sigma, \boldsymbol{X}_i &\overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma^2), \text{ for } i = 1, \ldots, n, \\ \sigma &\sim \text{InverseGamma}(1, 1), \\ p(\boldsymbol{\beta}) &\propto 1. \end{aligned} \tag{14}$$

The robust Bayesian mean regression model used the SNCP likelihood (Arellano-Valle and Azzalini, 2008), i.e.,

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \sigma, \boldsymbol{X}_i &\overset{\text{ind}}{\sim} \text{SNCP}(\boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma^2, \gamma_1), \text{ for } i = 1, \ldots, n, \\ \sigma &\sim \text{InverseGamma}(1, 1), \\ \gamma_1 &\sim \text{Uniform}(-1, 1), \\ p(\boldsymbol{\beta}) &\propto 1. \end{aligned} \tag{15}$$

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

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \sigma, \boldsymbol{X}_i &\overset{\text{ind}}{\sim} \text{ALD}(\boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma,\, p = 0.5), \text{ for } i = 1, \ldots, n, \\ \sigma &\sim \text{InverseGamma}(1, 1), \\ p(\boldsymbol{\beta}) &\propto 1. \end{aligned} \tag{16}$$

We stress that in this simulation study, *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.

With two different sample sizes $n = 30$ and $n = 300$, we generated observations from the model,

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

where $\beta_0 = \beta_1 = 1$ and, for $i = 1, \ldots, n$, $\epsilon_i \overset{\text{i.i.d}}{\sim} 0.05\mathcal{N}(-50, 1^2) + 0.95\mathcal{N}(0, 1^2)$ 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 (9), the DTP-Student-t model (10), and the TPSC-Student-t model (11). 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, SNCP, ALD, and TPSC-Student-t likelihoods.

In Figure 4, 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. 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 replicate.

Figure 5 depicts the true mode-zero model error distribution contrasted with the four estimated mode-zero error distributions based on the regression analyses considered in Figure 4, where we used posterior means as point estimates for all parameters. For mean regression analysis, the estimated normal distribution (with $\hat{\sigma} = 12.7$) and the estimated SNCP distribution (with $\hat{\sigma} = 13.3$ and $\hat{\gamma}_1 = 0.08$) both failed to capture the shape of the true distribution and indicated much greater variability around the mode than there truly was. The estimated error distribution from the median regression model (with $\hat{\sigma} = 2.13$ in the ALD likelihood) was much improved over the former two estimated densities. However, the estimation continued to improve notably when the TPSC-Student-t distribution was assumed in the proposed modal regression (with $\hat{w} = 0.55, \hat{\sigma} = 0.80, \hat{\delta} = 1.08$). In particular, the height of this last estimated density at the mode was the closest to that of the true error distribution. Compared with the three competitors, the tail behavior of the estimated TPSC-Student-t density was also strikingly similar to that of the ground truth.

We repeated this experiment comparing the four regression models for 300 times, with sample sizes of $n = 30$ and $n = 300$ in each replication. Table 2 provides Monte Carlo averages of the coverage rate of the 90% HPD prediction interval, width of the prediction interval, and ELPD for each regression model. When the sample size is as small as 30, the two mean regression models yielded lower coverage rates yet wider prediction intervals on average than those from the median and modal regression models. Even though the latter two regression models enjoyed similar (higher) coverage rates, the modal regression model tended to give much tighter prediction intervals.

Note that on average 95% of the data generated in this simulation study were non-outliers. Coverage rates close to 95% thus imply that these models have adequate prediction intervals as they can predict the values of non-outliers reasonably well. Within the methods with similar coverage rates, we prefer the modal regression model with the narrowest prediction interval since these predictions have the least amount of uncertainty. Furthermore, the modal regression model with the TPSC-Student-t likelihood had the largest average Expected Log Predictive Density (ELPD), reinforcing that the modal regression model based on the TPSC-Student-t provided the best overall

Table 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. The 90% HPD intervals were used to calculate the coverage rate.

Sample Size	Likelihood (regression model)	Coverage Rate (%)	Width	ELPD
n = 30	normal	93.47(0.20)	32.26(1.08)	-104.70(2.00)
	SNCP	93.56(0.19)	33.78(1.13)	-105.60(2.02)
	ALD	94.11(0.21)	15.47(0.56)	-85.11(1.40)
	TPSC-Student-t	**94.69(0.22)**	**8.31(0.21)**	**-59.96(0.64)**
n = 300	normal	95.01(0.07)	35.67(0.26)	-1142.65(3.05)
	SNCP	95.01(0.07)	35.85(0.27)	-1144.74(3.02)
	ALD	95.03(0.07)	14.94(0.16)	-861.39(3.45)
	TPSC-Student-t	**94.64(0.06)**	**6.69(0.06)**	**-591.81(1.87)**

model fit.

After we increased the sample size from $n = 30$ to $n = 300$, all regression methods had around 95% coverage rate. However, the modal regression model still had the narrowest prediction interval and highest ELPD, on average. This again confirms the supremacy of the modal regression model in this simulation study.

5.2. Right-skewed data

In Section 5.1, we demonstrated the advantages of our Bayesian modal regression models over Bayesian mean and median regression when the data was left-skewed. In this section, we compare the performance of our model on right-skewed data against an alternative modal linear regression (MODLR) method proposed by Yao and Li (2014). Whereas our Bayesian modal regression models are fully parametric, the mode-zero model error distribution under MODLR is left unspecified. However, just like our models, MODLR assumes a linear function of covariates for the conditional mode of Y, i.e. $\text{Mode}(Y \mid \boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{\beta}$.

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

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i,$$

where $\beta_0 = \beta_1 = \beta_2 = 1$, $X_{1,i}$ and $X_{2,i}$ come from the standard normal distribution independently, and ϵ_i follows $\text{SNDP}(-0.3754, 1, 5)$, which is

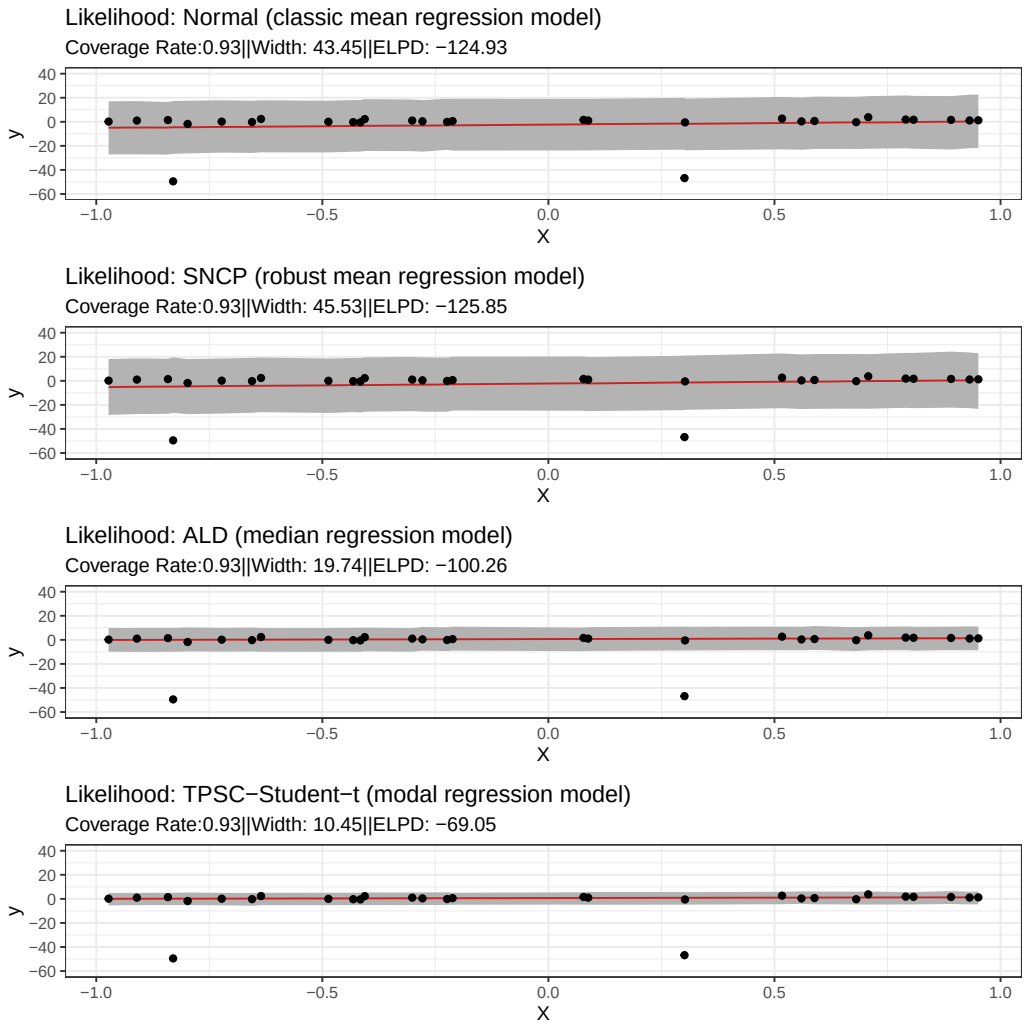

Figure 4: *The gray shaded areas show the* 90% *posterior prediction intervals for the simulated left-skewed data. The solid red line is the estimated median from the posterior predictive distribution. The prediction intervals are narrower for Bayesian modal regression.*

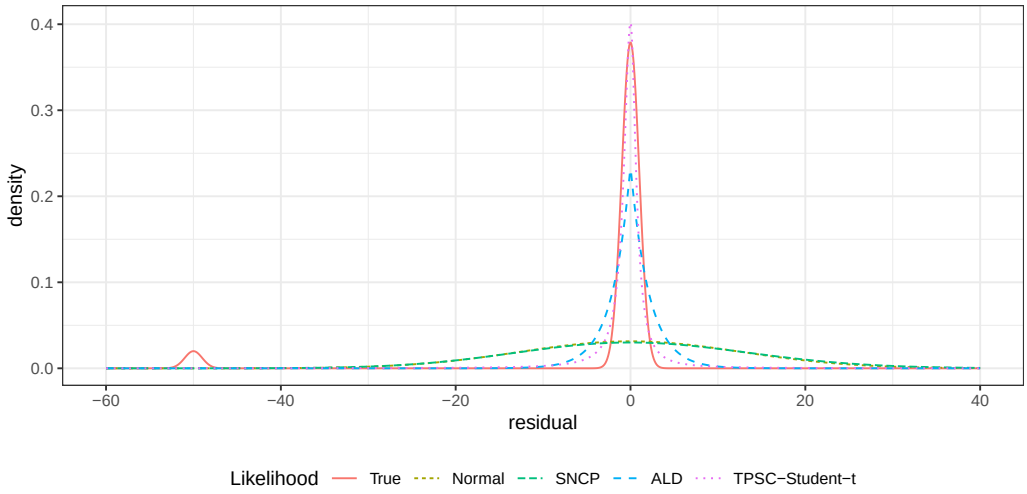

Figure 5: *The estimated density plots of residuals with fixed mode at* 0 *from regression models utilizing normal, SNCP, ALD and TPSC-Student-t likelihoods.*

an SNDP distribution with the location parameter as -0.3754, the scale parameter as 1, and the skewness parameter as 5 (Azzalini, 2013). This distribution is right-skewed and with mode at 0. We fit a Bayesian modal regression that assumes the TPSC-Student-t distribution of Y given X_1 and X_2 to the simulated data.

We also implemented MODLR to infer covariate effects. MODLR is based on kernel density estimation and requires bandwidth selection. To ensure a (more than) fair comparison, we selected the bandwidth that minimizes the average sum-of-square bias, $\sum_{j=0}^{2}(\hat{\beta}_j - \beta_j)^2$, across 300 repeated experiments. This bandwidth selection procedure is biased towards MODLR as it uses the true values of regression coefficients that are unknown in practice. After the optimal bandwidth was chosen, we used MODLR to estimate regression coefficients and constructed confidence intervals using the nonparametric bootstrap with 500 bootstrap samples. We repeated this process for 300 total experiments and compared the results to our Bayesian modal regression model. As in Section 2.2, drawing inference for regression coefficients based on our parametric modal regression model was more straightforward, with the posterior means serving as the point estimates and the corresponding credible intervals as interval estimates.

Table 3 presents summary statistics for inferences of the regression coeffi-

Table 3: *Comparison of the Bayesian modal regression based on the TPSC-Student-t and MODLR. 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. CI denotes 90% credible interval.*

Regression Model	Parameter	Point Estimation	Coverage Rate (%)	Width of CI
TPSC-Student-t	β_0	1.18 (0.17)	91.67	0.76 (0.18)
	β_1	1.00 (0.12)	91.33	0.41 (0.09)
	β_2	1.00 (0.12)	92.00	0.41 (0.09)
MODLR	β_0	1.25 (0.15)	42.00	0.54 (0.18)
	β_1	1.00 (0.16)	89.67	0.55 (0.22)
	β_2	1.01 (0.16)	91.33	0.56 (0.23)

cents under the two considered methods averaged across 300 Monte Carlo replicates. Both Bayesian modal regression based on TPSC-Student-t and MODLR produced point estimates for the covariate effects β_1 and β_2 close to their true values, with the TPSC-Student-t being more precise than MODLR. Both methods appeared to overestimate the intercept, possibly more so when MODLR was used. Table 3 shows that the interval estimates from the proposed Bayesian modal regression model were clearly more reliable than those from MODLR, especially those for β_1 and β_2 where our Bayesian model yielded tighter interval estimates with similar empirical coverage rates. In this example, Bayesian modal regression based on TPSC-Student-t exhibited superior performance over MODLR in terms of inference for regression coefficients. This is despite the fact that the likelihood in our parametric Bayesian model is misspecified and despite the fact that MODLR is ostensibly more flexible by allowing the residual error distribution to be unspecified.

6. More data applications of Bayesian modal regression

6.1. Boston housing prices

To demonstrate the differences in interpretation of mean/median/modal regression, we analyzed a dataset containing $n = 506$ house prices in the area of Boston, Massachusetts, in the year 1970. This dataset is available in the R package `MASS` (Venables and Ripley, 2013) under the name `Boston`. We

are interested in the intricate relationship between house prices and thirteen covariates recorded in the dataset. The models are formulated as follows:

$$\mathbb{M}\left(Y_i \mid \boldsymbol{\beta}\right) = \boldsymbol{X}_i^\top \boldsymbol{\beta}, \text{ for } i = 1, \ldots, 506,$$

where the response Y_i is the median value of owner-occupied homes in thousands dollars in the ith area, and $\boldsymbol{X}_i$ contains a 1 (for the intercept β_0) and values for the following 13 covariates associated with the same ith area: per capita crime rate by town ($X_{1,i}$), proportion of residential land zoned for lots over 25,000 square feet ($X_{2,i}$), proportion of non-retail business acres per town ($X_{3,i}$), Charles River dummy variable ($X_{4,i} = 1$ if tract bounds river; 0 otherwise), Nitrogen oxides concentration in parts per 10 million ($X_{5,i}$), average number of rooms per dwelling ($X_{6,i}$), proportion of owner-occupied units built prior to 1940 ($X_{7,i}$), weighted mean of distances to five Boston employment centres ($X_{8,i}$), an index of accessibility to radial highways ($X_{9,i}$), full-value property-tax rate per \$10,000 ($X_{10,i}$), pupil-teacher ratio by town ($X_{11,i}$), the modified proportion of blacks by town ($X_{12,i}$), and percentage of the population that was lower status ($X_{13,i}$).

Parallel to the study design in Section 5.1, we carried out four different regression analyses of this dataset, including the usual mean regression with normal likelihood, the more robust mean regression assuming a skewed normal model error (SNCP), median regression based on the ALD likelihood, and our proposed modal regression assuming a TPSC-Student-t distribution for the model error distribution. Table 4 reports estimates of the fourteen regression coefficients resulting from these four different analyses.

Because the interpretation of regression coefficients depends on whether the regression function is the conditional mean, median, or mode of the response, we focus on comparing conclusions from different regression models with respect to statistical significance of a covariate effect on the house price. According to Table 4, all four regression analyses reached the same conclusions for all covariates in this regard except for one covariate, the proportion of owner-occupied units built prior to 1940 (see results for β_7 in Table 4). In particular, the 90% credible intervals from the two mean regression models both contained zero, suggesting that, on average, the house price in Boston is not significantly influenced by the proportion of owner-occupied units built prior to 1940. Yet the counterpart results from the median regression model and the modal regression indicate that, had one looked into the association of the considered covariates with the median or the mode of house prices, one

Table 4: Parameter estimates obtained from the mean (normal and SNCP)/median (ALD)/modal regression models (TPSC-Student-t) fitted to the Boston house price data. The mean, 5% quantile, and 95% quantile of the posterior distribution of each regression coefficient are listed under Mean, q5, and q95, respectively.

	Normal			SNCP			ALD			TPSC-Student-t		
Parameter	Mean	q5	q95	Mean	q5	q95	Mean	q5	q95	Mean	q5	q95
β_0	36.46	28.12	44.81	35.73	28.60	42.90	14.91	7.32	22.48	13.02	6.48	19.66
β_1	-0.11	-0.16	-0.05	-0.12	-0.17	-0.07	-0.12	-0.16	-0.06	-0.13	-0.17	-0.09
β_2	0.05	0.02	0.07	0.03	0.01	0.05	0.04	0.02	0.05	0.02	0.01	0.04
β_3	0.02	-0.08	0.12	0.03	-0.05	0.11	0.01	-0.05	0.07	0.01	-0.05	0.06
β_4	2.69	1.27	4.11	1.38	0.17	2.54	1.49	0.51	2.53	1.44	0.50	2.39
β_5	-17.80	-24.06	-11.53	-14.02	-19.48	-8.69	-8.99	-13.64	-4.36	-6.71	-10.73	-2.79
β_6	3.81	3.12	4.49	3.42	2.77	4.08	5.26	4.50	6.01	4.87	4.11	5.63
β_7	0.00	-0.02	0.02	-0.02	-0.03	0.00	-0.03	-0.04	-0.01	-0.04	-0.05	-0.02
β_8	-1.48	-1.81	-1.15	-1.12	-1.42	-0.83	-0.99	-1.23	-0.75	-0.89	-1.11	-0.66
β_9	0.31	0.20	0.42	0.24	0.14	0.34	0.18	0.09	0.26	0.14	0.07	0.21
β_{10}	-0.01	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
β_{11}	-0.95	-1.17	-0.74	-0.82	-1.00	-0.64	-0.75	-0.90	-0.59	-0.61	-0.73	-0.48
β_{12}	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.02	0.01	0.01	0.01
β_{13}	-0.52	-0.61	-0.44	-0.43	-0.51	-0.35	-0.31	-0.39	-0.23	-0.28	-0.35	-0.21

would conclude that the proportion of owner-occupied units built prior to 1940 has a significant, negative effect on house price.

To provide further context and insights, we present the posterior prediction coverage rate of the Boston house prices, the width of prediction interval, and ELPD in Table 5. Using our proposed modal regression, the prediction intervals attained similar empirical coverage rates, while also giving the tightest prediction intervals on average among the four models considered. Our modal regression model also provided the overall best fit for this dataset according to ELPD.

Table 5: Comparison of Bayesian mean/median/modal regression models fitted to the Boston house price data.

Likelihood (regression model)	Coverage Rate (%)	Width	ELPD
Normal (mean regression)	93.28	15.82	-1518.66
SNCP (mean regression)	91.30	14.38	-1464.62
ALD (median regression)	90.51	14.55	-1444.57
TPSC-Student-t (modal regression)	89.33	13.12	-1408.28

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 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 6 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 was 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 6 shows that the ELPD of the modal regression model based on the FG likelihood (9) was 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.

7. Discussion

In this paper, 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. 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. Our framework adds to the existing literature on likelihood-based robust regression (Ronchetti and Huber, 2009; Box and Tiao, 1968; Lange et al.,

Table 6: 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)	-627.12	β_0 (intercept)	3.09	2.46	3.73
		β_1 (Age)	0.96	0.44	1.47
		β_2 $\left(\text{Age}^2\right)$	-0.05	-0.13	0.04
ALD (median regression)	-638.12	β_0 (intercept)	2.81	2.12	3.55
		β_1 (Age)	1.12	0.54	1.69
		β_2 $\left(\text{Age}^2\right)$	-0.07	-0.16	0.03
FG (modal regression)	-623.18	β_0 (intercept)	2.37	1.85	2.89
		β_1 (Age)	1.15	0.72	1.59
		β_2 $\left(\text{Age}^2\right)$	-0.11	-0.18	-0.03

1989; da Silva et al., 2020; Gagnon et al., 2020). In particular, the GUD family exhibits robustness to outliers, skewness, and model misspecification, as demonstrated in Section 5. In contrast to other parametric families designed for robust regression (Azzalini, 2013), however, a notable feature of the GUD family is that all members of this family have a location parameter that is *also* the conditional mode. This makes the GUD family suitable for inference and prediction of the conditional mode specifically.

Compared to mean and quantile regression, work on Bayesian modal regression analysis is quite scarce. Our paper 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 2.1 and the Boston house prices data in Section 6.1), criminology (the murder rate data in Section 2.2), and molecular biology (the serum data in Section 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.

We stress that it is important not to fit only one type of regression model. In practice, researchers should fit several models to the data and utilize regression diagnostics to evaluate model assumptions and determine whether there are any influential observations. Our modal regression model framework is an especially appealing choice when the data is skewed and(or) contains (extreme) outliers. Moreover, model selection criteria such as ELPD can be used to select a suitable (mean, quantile, or modal) regression model for final analyses. Other posterior predictive checks, e.g. those described in Chapter 6 of Gelman et al. (2013), can also be used to assess the appropriateness of using a GUD likelihood for Bayesian inference.

The modal regression models in this paper 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 manuscript, 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.

Another interesting future direction to explore is Bayesian modal regression in high dimensions. When the number of covariates p is large or even exceeds the sample size n, some form of regularization is typically desired. In this case, we can replace the flat prior on $\boldsymbol{\beta}$ with a spike-and-slab prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Ročková and George, 2018) or a global-local shrinkage prior (Bhadra et al., 2019; Griffin and Brown, 2021). These priors shrink most of the regression coefficients in $\boldsymbol{\beta}$ towards zero and allow for variable selection. We anticipate that these types of priors would work well in high-dimensional Bayesian modal regression with the GUD likelihood, especially if $p > n$.

Declaration of generative AI in scientific writing

During the preparation of this work the authors used ChatGPT in order to check grammar. After using this tool/service, the authors reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgments

We are grateful to the Associate Editor and two anonymous reviewers for their thoughtful comments and suggestions which helped to greatly improve our article. The last listed author was partially support by National Science Foundation grant DMS-2015528.

References

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

Arellano-Valle, R.B., Azzalini, A., 2008. The centred parametrization for the multivariate skew-normal distribution. Journal of Multivariate Analysis 99, 1362–1382.

Aristodemou, K., 2014. New regression methods for measures of central tendency. Ph.D. thesis. Brunel University.

Azzalini, A., 2013. The Skew-Normal and Related Families. Cambridge University Press.

Behboodian, J., 1970. On the modes of a mixture of two normal distributions. Technometrics 12, 131–139.

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

Bhadra, A., Datta, J., Polson, N.G., Willard, B., 2019. Lasso meets horseshoe: A survey. Statistical Science 34, 405 – 427.

Boos, D.D., Stefanski, L.A., 2013. Essential Statistical Inference: Theory and Methods. Springer.

Bourguignon, M., Leão, J., Gallardo, D.I., 2020. Parametric modal regression with varying precision. Biometrical Journal 62, 202–220.

van de Bovenkamp, F.S., Hafkenscheid, L., Rispens, T., Rombouts, Y., 2016. The emerging importance of IgG Fab glycosylation in immunity. The Journal of Immunology 196, 1435–1441.

Box, G.E., Tiao, G.C., 1968. A Bayesian approach to some outlier problems. Biometrika 55, 119–129.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. Journal of Statistical Software 76, 1–32.

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

Chen, Y.C., Genovese, C.R., Tibshirani, R.J., Wasserman, L., 2016. Nonparametric modal regression. The Annals of Statistics 44, 489–514.

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

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

Durante, D., 2019. Conjugate Bayes for probit regression via unified skew-normal distributions. Biometrika 106, 765–779.

Ekawati, D., Warsono, W., Kurniasari, D., 2015. On the moments, cumulants, and characteristic function of the log-logistic distribution. IPTEK The Journal for Technology and Science 25, 78–82.

Fernández, C., Steel, M.F., 1998. On Bayesian modeling of fat tails and skewness. Journal of the American Statistical Association 93, 359–371.

Gagnon, P., Desgagné, A., Bédard, M., 2020. A new Bayesian approach to robustness against outliers in linear regression. Bayesian Analysis 15, 389–414.

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

George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88, 881–889.

Geweke, J., 1993. Bayesian treatment of the independent Student-t linear model. Journal of Applied Econometrics 8, S19–S40.

Griffin, J.E., Brown, P.J., 2021. Bayesian global-local shrinkage methods for regularisation in the high dimension linear model. Chemometrics and Intelligent Laboratory Systems 210, 104255.

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

Ho, C.S., Damien, P., Walker, S., 2017. Bayesian mode regression using mixtures of triangular densities. Journal of Econometrics 197, 273–283.

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

Isaacs, D., Altman, D., Tidmarsh, C., Valman, H., Webster, A., 1983. Serum immunoglobulin concentrations in preschool children measured by laser nephelometry: Reference ranges for IgG, IgA, IgM. Journal of Clinical Pathology 36, 1193–1196.

Koenker, R., Chernozhukov, V., He, X., Peng, L., 2017. Handbook of Quantile Regression. Chapman and Hall/CRC.

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

Lange, K.L., Little, R.J., Taylor, J.M., 1989. Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84, 881–896.

Lee, M.j., 1989. Mode regression. Journal of Econometrics 42, 337–349.

Lee, M.J., 1993. Quadratic mode regression. Journal of Econometrics 57, 1–19.

Liu, Q., Huang, X., Zhou, H., 2024. The flexible Gumbel distribution: A new model for inference about the mode. Stats 7, 317–332.

Lochner, L., 2020. Chapter 9 - Education and crime, in: Bradley, S., Green, C. (Eds.), The Economics of Education (Second Edition). Academic Press, pp. 109–117.

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

Meredith, M., Kruschke, J., Meredith, M.M., 2018. Package 'hdinterval'. Highest (Posterior) Density Intervals URL: https://cran.r-project.org/web/packages/HDInterval/index.html.

Mitchell, T., Beauchamp, J., 1988. Bayesian variable selection in linear regression. Journal of the American Statistical Association 83, 1023–1032.

Neal, R.M., 2011. MCMC using Hamiltonian dynamics, in: Brooks, S., Gelman, A., Jones, G., Meng, X.L. (Eds.), Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC, pp. 113–162.

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

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods. Springer New York.

Ronchetti, E.M., Huber, P.J., 2009. Robust Statistics. Wiley.

Ročková, V., George, E.I., 2018. The spike-and-slab lasso. Journal of the American Statistical Association 113, 431–444.

Royston, P., 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, 429–453.

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

Sager, T.W., Thisted, R.A., 1982. Maximum likelihood estimation of isotonic modal regression. The Annals of Statistics 10, 690–707.

Shin, J.Y., Chen, S., Kim, T.W., 2015. 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, 366–375.

Siegel, A.F., 2016. Practical Business Statistics. Academic Press.

da Silva, N.B., Prates, M.O., Gonçalves, F.B., 2020. Bayesian linear regression models with flexible error distributions. Journal of Statistical Computation and Simulation 90, 2571–2591.

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

Smith, R.L., 2003. Statistics of extremes, with applications in environment, insurance, and finance, in: Finkenstadt, B., Rootzen, H. (Eds.), Extreme Values in Finance, Telecommunications, and the Environment. Chapman and Hall/CRC, pp. 20–97.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B: Statistical Methodology 64, 583–639.

Teicher, H., 1963. Identifiability of finite mixtures. The Annals of Mathematical Statistics 34, 1265–1269.

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27, 1413–1432.

Venables, W.N., Ripley, B.D., 2013. Modern Applied Statistics with S-PLUS. Springer.

Vidal, I., 2014. A Bayesian analysis of the Gumbel distribution: An application to extreme rainfall data. Stochastic Environmental Research and Risk Assessment 28, 571–582.

Watanabe, S., Opper, M., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 11, 3571–3594.

Xiang, S., Yao, W., 2022. Modal regression for skewed, truncated, or contaminated data with outliers, in: He, W., Wang, L., Chen, J., Lin, C.D. (Eds.), Advances and Innovations in Statistics and Data Science. Springer, pp. 257–273.

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

Yu, K., Aristodemou, K., 2012. Bayesian mode regression. arXiv preprint arXiv:1208.0579 .

Yu, K., Moyeed, R.A., 2001. Bayesian quantile regression. Statistics & Probability Letters 54, 437–447.

Yu, K., Zhang, J., 2005. A three-parameter asymmetric Laplace distribution and its extension. Communications in Statistics—Theory and Methods 34, 1867–1879.

Zhou, H., Huang, X., 2020. Parametric mode regression for bounded responses. Biometrical Journal 62, 1791–1809.

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