Log-Regularly Varying Scale Mixture of Normals for Robust Regression

Yasuyuki Hamura[a,*], Kaoru Irie[b], Shonosuke Sugasawa[c]

[a] *Graduate School of Economics, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, JAPAN. JSPS Research Fellow.*
[b] *Faculty of Economics, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, JAPAN.*
[c] *Center for Spatial Information Science, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba, JAPAN.*

Abstract

Linear regression that employs the assumption of normality for the error distribution may lead to an undesirable posterior inference of regression coefficients due to potential outliers. A finite mixture of two components, one with thin and one with heavy tails, is considered as the error distribution in this study. For the heavily-tailed component, the novel class of distributions is introduced; their densities are log-regularly varying and have heavier tails than the Cauchy distribution. Yet, they are expressed as a scale mixture of normals which enables the efficient posterior inference when using a Gibbs sampler. The robustness of the posterior distributions is proved under the proposed models using a minimal set of assumptions, which justifies the use of shrinkage priors with unbounded densities for the coefficient vector in the presence of outliers. An extensive comparison with the existing methods via simulation study shows the improved performance of the proposed model in point and interval estimation, as well as its computational efficiency. Further, the posterior robustness of the proposed method is confirmed in an empirical study with shrinkage priors for regression coefficients.

Keywords: Robust statistics, Linear regression, Heavily-tailed distribution, Scale mixture of normals, Log-regularly varying density, Gibbs sampler

*Corresponding author

Email addresses: yasu.stat@gmail.com (Yasuyuki Hamura), irie@e.u-tokyo.ac.jp (Kaoru Irie), sugasawa@csis.u-tokyo.ac.jp (Shonosuke Sugasawa)

1. Introduction

The robustness of outliers in linear regression models has been studied extensively for its importance, and the research on theory and methodology for robust statistics has been accumulated in the past years. In the full posterior inference, the concept of robustness is not limited to point estimation but targets the posterior distributions of parameters of interest. Also known as outlier-proneness or outlier-rejection, posterior robustness defines the property of posterior distributions in which the difference of posteriors with and without outliers diminishes as the values of outliers become extreme (O'Hagan, 1979). A series of studies on posterior robustness has revealed sufficient conditions for error distributions in order to achieve robustness and has provided the specific error distributions that meet such conditions (see the detailed review by O'Hagan and Pericchi 2012). Specifically, the error distribution must be sufficiently heavily-tailed to account for outliers (Andrade and O'Hagan, 2006, 2011). Recent studies have identified the necessity of log-regularly varying error density functions for the full posterior robustness (Desgagné, 2015; Desgagné and Gagnon, 2019). In the context of linear models, Gagnon et al. (2019) show that full robustness is obtained if the error distribution is log-regularly varying and proposed the use of the log-Pareto-tailed normal (LPTN) distribution as a log-regularly varying error distribution that has been practically non-existent. Therefore, the main objective of this study is to propose a new log-regularly varying alternative to the LPTN distribution. The robustness concept of our interest in this study is limited to the full posterior robustness (See Theorem 1).

In contrast to the truncation approach, we model the error distribution using a finite mixture of two components: one with thinner tails, such as normal distributions, and the other with super heavy tails to accommodate potential outliers (Box and Tiao, 1968). This simple, intuitive approach to modeling outliers has received less attention in the methodological literature but is routinely practiced in applied statistics (see Carter and Kohn 1994; West 1997; Frühwirth-Schnatter 2006; Tak et al. 2019; Silva et al. 2020). The heavy-tailed component of the mixture remains in the general class of a scale mixture of normals (West, 1984), which allows conditional conjugacy for efficient posterior computation. The aforementioned LPTN distribution can also be cast as the two-component mixture (Desgagné, 2021), but does

not allow the scale mixture representation. In this study, we propose a super heavy-tailed distribution that is represented as a scale mixture of normals.

For the super heavy-tailed distribution that comprises the finite mixture, the Student's t-distribution is still regarded as thin-tailed for its outlier sensitivity. We propose the use of distributions that have been utilized in the robust inference for high-dimensional count data (Hamura et al., 2019) for their super heavy tails. This is the log-Pareto mixture of normals (LPMN), and has another mixture representation by using the gamma distribution with the hierarchical structure on shape parameters, which enables posterior inference by a simple but efficient Gibbs sampler. The tails of these distributions are heavier than those of Cauchy distributions. In fact, the density of the proposed error distribution is log-regularly varying, similar to those of other super heavy-tailed distributions considered for posterior robustness, including LPTN distributions.

The proposed error distribution is the finite mixture of the standard normal and LPMN distributions, or the N-LPNM distribution for short. We provide a set of sufficient conditions for the posterior robustness under the linear regression models with the N-LPMN distribution, that is different from the conditions used in Gagnon et al. (2019). The conditions we use in proving the posterior robustness restrict the available priors for the regression coefficients and observational scale, but do not exclude the use of unbounded prior densities, including shrinkage priors (e.g., horseshoe priors, Carvalho et al. 2009, 2010). As a result, the robustness under shrinkage, or variable selection, is within the scope of our research. In empirical studies, we practice robust posterior inference for linear regression models with the horseshoe prior for illustration.

The reminder of this paper is organized as follows. In Section 2, we introduce the new error distribution and describe its use in linear regression models, followed by the theoretical results on posterior robustness. The algorithm for posterior computation is provided in Section 3, followed by a discussion on its computational efficiency. In Section 4, we conduct out simulation studies to compare the proposed method to existing models. In Section 5, we illustrate the proposed method using two famous datasets: the Boston housing data and diabetes data. The paper is concluded, with a discussion on future works, in Section 6. The R code implementing the proposed method is available at the GitHub repository (`https://github.com/sshonosuke/EHE`).

2. A new error distribution for robust regression

2.1. Linear models and error distributions

Let y_i be a response variable and x_i be an associated p-dimensional vector of covariates, for $i = 1, \ldots, n$. We consider a linear regression model, $y_i = x_i^\top \beta + \sigma\varepsilon_i$, where β is a p-dimensional vector of regression coefficients and σ is an unknown scale parameter. The error terms, $\varepsilon_1, \ldots, \varepsilon_n$, are directly linked to the posterior robustness. Modeling these errors simply by Gaussian distributions makes the posterior inference very sensitive to outliers.

To define the error distribution, we introduce a latent variable, u_i, and assume that the error distribution is conditionally Gaussian, as $\varepsilon_i | u_i \sim N(0, u_i)$. A typical choice of the distribution of u_i is the inverse-gamma distribution, which leads to the marginal distribution of ε_i being the t-distribution. However, as shown in Gagnon et al. (2019) and in our main theorem, this choice does not hold the desirable robustness properties for posterior distributions, even when the distribution of ε_i is a Cauchy distribution.

As an error distribution whose density function is log-regularly varying, Gagnon et al. (2019) proposes the LPTN distribution, which replaces the thin tails of the standard normal distribution by the super heavy tails of a log-Pareto distribution. Despite its desirable robustness, this truncation approach complicates the likelihood function, making the posterior inference under the LPTN error distribution challenging. Several parameters, including the regression coefficients, cannot be directly sampled from the conditional posteriors. In addition, the class of LPTN distributions has a hyperparameter, for which a conditionally conjugate prior is not available. These challenges may require the use of the Metropolis-Hastings algorithm and lead to an increased computational cost under the LPTN models. Consequently, the LPTN distribution is not readily available in the context of more structured linear models with random effects.

2.2. Log-Pareto mixture of normals

As stated in the introduction, the error distribution in this study is not a single continuous mixture of normals. Instead, it is a mixture of two components. We introduce latent binary variable z_i and model it using $\Pr[z_i = 1] = 1 - \Pr[z_i = 0] = s$ with weight $s \in (0, 1)$. If $z_i = 0$, then the error distribution is simply the standard normal distribution, that is, $\epsilon_i | (u_i, z_i = 0) \sim N(0, 1)$. If $z_i = 1$, then we consider the scale mixture of normals with latent variable u_i as $\varepsilon_i | (u_i, z_i = 1) \sim N(0, u_i)$, where u_i follows

some (super) heavy-tailed distribution on $(0, \infty)$ with density H. Preparing two distributions in modeling of the error distribution is based on the work of Box and Tiao (1968); the first component generates non-outlying errors and the second component is intended to absorb outlying errors. We require the mixing distribution of u_i to be log-regularly varying (Desgagné, 2015), or $H(u; \gamma) \approx u^{-1}(\log u)^{-1-\gamma}$ as $u \to \infty$ with some parameter $\gamma > 0$.

For the H-distribution, we consider a log-Pareto distribution whose density is given by

$$H(u; \gamma) = \frac{\gamma}{1+u} \frac{1}{\{1 + \log(1+u)\}^{1+\gamma}}, \quad u > 0. \tag{1}$$

The log-Pareto distribution can be obtained from the Pareto distribution via change of variables and has some variations. The version (1) is found in Cormann and Reiss (2009). Under this choice for H, the super heavy-tailed component of the finite mixture is:

$$f_{\text{LPMN}}(\varepsilon_i; \gamma) = \int_0^\infty \phi(\varepsilon_i; 0, u_i) H(u_i; \gamma) du_i, \tag{2}$$

and the marginal distribution of ε_i is obtained as:

$$f_{\text{N-LPMN}}(\varepsilon_i; s, \gamma) = (1-s)\phi(\varepsilon_i; 0, 1) + s f_{\text{LPMN}}(\varepsilon_i; \gamma), \tag{3}$$

where $\phi(\varepsilon_i; 0, u)$ is the normal density with a mean of zero and variance u. The new error distribution in (3) is the mixture of the standard normal and LPMN distributions, or the N-LPMN distribution. The second component, or the LPMN distribution, is a scale mixture of normals but does not admit any closed form expression. To handle this component in posterior computation, as we see later in Section 3.1, we utilize the augmentation of the H-distribution using several gamma-distributed state variables. Through this augmentation, the posterior inference for this model becomes straightforward.

A notable property of this new error distribution is its super heavy tails, as shown in the following proposition. The relevant proof is presented in the Supplementary Materials.

Proposition 1. *The densities (1), (2) and (3) satisfy,*

$$H(|x|; \gamma) \approx f_{\text{LPMN}}(x; \gamma) \approx f_{\text{N-LPMN}}(x; s, \gamma) \approx |x|^{-1}(\log |x|)^{-1-\gamma},$$

for large $|x|$, and for any $\gamma > 0$ and $s > 0$.

The above proposition shows that the super heavy tail of the H-distribution is inherited to the LPMN distribution, then to the N-LPMN distribution. As a result, the density of the N-LPMN distribution belongs to a family of log-regularly varying functions. Notably, the tails of the N-LPMN density are heavier than those of the Cauchy distribution, $f_C(x) \approx |x|^{-2}$.

Figure 1 shows the cumulative distribution functions (CDFs) of the H and LPMN distributions for $\gamma = 0.5, 1$ and 2. The tails of LPMN distributions are heavier than those of the Cauchy distribution, as shown in the right panel. This fact is also confirmed by comparing the CDFs of the H and inverse-gamma distributions in the left panel. The property of the super heavy tails of the H and LPMN densities leads to posterior robustness in Theorem 1. The density function of the N-LPMN distribution in (3) is plotted on Figure 2 for $s = 0.05, 0.1$ and 0.2. We observe that the shape of the N-LPMN distribution is similar to the standard normal distribution around the origin. The tails become heavier as the mixture weight s increases.

2.3. Definition of outliers

We first specify the structure of the outliers, based on the definition by Desgagné and Gagnon (2019). The set of indices for n observations, $\{1, \ldots, n\}$, is split into two disjoint subsets, $\mathcal{K}$ and $\mathcal{L}$, which represent those of the non-outlying and outlying values, respectively. Note that $\mathcal{K} \cup \mathcal{L} = \{1, \ldots, n\}$ and $\mathcal{K} \cap \mathcal{L} = \emptyset$. Let $\mathcal{D} = \{y_1, \ldots, y_n\}$ be the set of observed data. The set of non-outlying observations is defined by $\mathcal{D}^* = \{y_i | i \in \mathcal{K}\}$.

The concept of (non-)outliers is defined by the observed values specified as,

$$y_i = \begin{cases} a_i, & \text{if } i \in \mathcal{K}, \\ a_i + b_i\omega, & \text{if } i \in \mathcal{L}, \end{cases}$$

where $a_i \in \mathbb{R}$, $b_i \neq 0$ and $\omega > 0$. We assume that ω is sufficiently large, such that the value of y_i for $i \in \mathcal{L}$ becomes extremely large, either positively or negatively. We define posterior robustness as the limiting behavior of the posterior of (β, σ^2) when ω tends to infinity. That is, the model is robust if the two posteriors, one of which is conditioned by the full dataset $\mathcal{D}$ and the other of which is conditioned by the dataset without the outliers $\mathcal{D}^*$, are equivalent when $\omega \to \infty$. In other words, when considering posterior

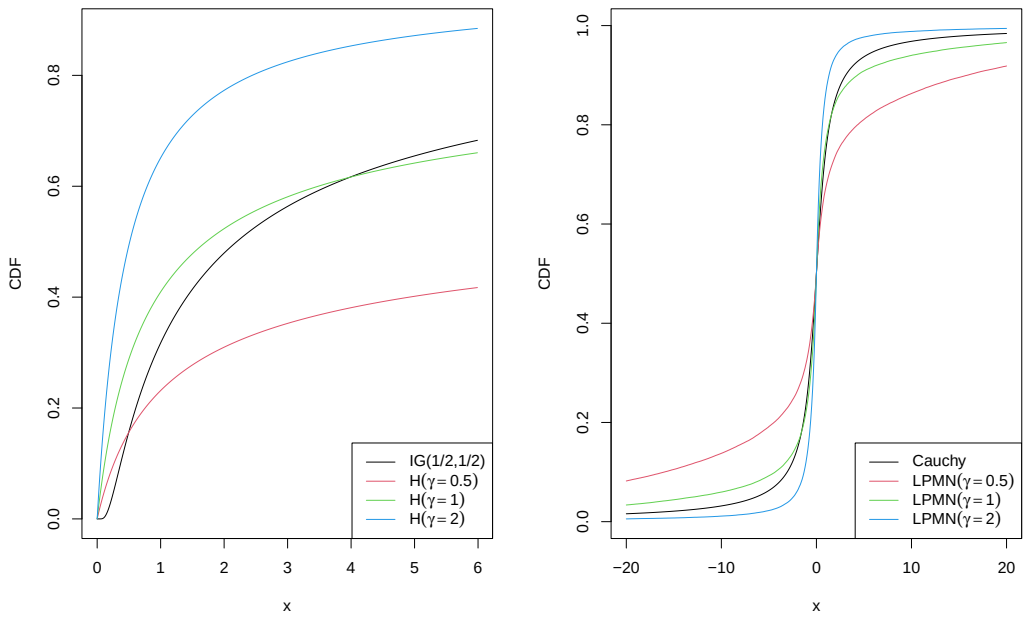

Figure 1: (Left): Cumulative distribution functions of $H(u;\gamma)$ for $\gamma \in \{0.5, 1.0, 2.0\}$, and the inverse gamma distribution with shape and scale 0.5. (Right): The empirical cumulative distributions of the LPMN distributions with $\gamma \in \{0.5, 1.0, 2.0\}$, using the Monte Carlo integration and compared with the distribution function of Cauchy distribution.

robustness, the outlying values are automatically discarded in the posterior inference without knowing which observations are outliers.

2.4. Robustness for the N-LPMN distribution

The class of prior distributions for (β, σ) for which we prove the posterior robustness is, for $k = 1, \ldots, p$,

$$\beta_k|\sigma \sim \frac{1}{\sigma}\pi_\beta\Big(\frac{\beta_k}{\sigma}\Big) \quad \text{and} \quad \sigma \sim \pi_\sigma(\sigma), \tag{4}$$

where $\beta_1, \ldots, \beta_p$ are conditionally independent given σ, and π_β and π_σ are the probability density functions on $\mathbb{R}$ and $(0, \infty)$, respectively. We limit our focus to the class of proper priors. This is because improper priors, such as the constant prior for β, have already been considered in Gagnon et al. (2019). Let $p(\beta, \sigma|\mathcal{D})$ be the posterior distribution of (β, σ) under the

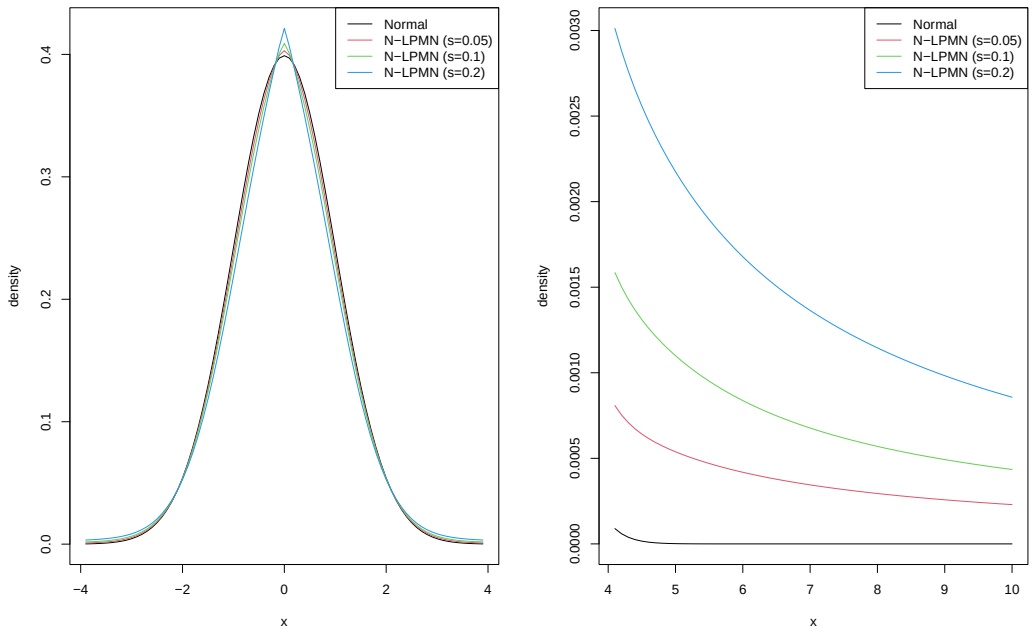

Figure 2: Densities of the proposed error distribution with $\gamma = 1$, $s \in \{0.05, 0.1, 0.2\}$, and standard normal error distribution. The intractable integral of the second component is computed using the Monte Carlo integration.

linear regression model with the N-LPMN distribution. Under this prior, the following theorem provides sufficient conditions for the posterior with the outliers converging toward that without the outliers as $\omega \to \infty$. Our proof is provided in the Supplementary Materials.

Theorem 1. *Assume that there exists* $c > 0$ *such that*

(A.1) $|\mathcal{K}| \geq |\mathcal{L}| + p$, *i.e., #non-outliers* $\geq$ *#outliers* $+$ *#predictors,*

(A.2) $\sup_{t \in \mathbb{R}} \{|t|^c \pi_\beta(t)\} < \infty$, *and*

(A.3) the prior moments of $\sigma^{-|\mathcal{K}|}$, σ^{c-1}, *and* σ^{c-n} *are all finite.*

Then the linear regression model with the error distribution in (3) and the prior in (4) is posterior robust, that is,

$$\lim_{\omega \to \infty} p(\beta, \sigma | \mathcal{D}) = p(\beta, \sigma | \mathcal{D}^*)$$

for all $(\beta, \sigma) \in \mathbb{R}^p \times (0, \infty)$.

The three assumptions above are met in many examples encountered in practice. Assumption (A.1) is the requirement for the number of non-outlying observations to be sufficiently large. Similar assumptions can be found in the literature (e.g., Theorem 2.1 (ii), Gagnon et al. 2019); however, (A.1) is of a simpler form and less restrictive. Frequently, many non-outlying observations comprise the majority of the dataset, thereby satisfying Assumption (A.1).

Assumption (A.2) limits the choice of priors for β, but still covers the wide class of probability distributions. For example, this assumption is always satisfied when $\pi_\beta(t)$ is bounded and $O(1/|t|)$ as $|t| \to \infty$. Examples of such a prior include the normal and t-distributions. However, note that (A.2) does not force the prior density π_β to be bounded, unlike the settings of Gagnon et al. (2019). An important example of this is the horseshoe prior, whose density is unbounded at the origin (Theorem 1, Carvalho et al. 2010). Subsequently, the horseshoe prior satisfies (A.2) for any $c \in (0, 2]$. As evident in this example, Theorem 1 can be a useful device to check the posterior robustness for the broader class of statistical problems, including the variable selection by the shrinkage priors.

Assumption (A.3) is the moment conditions for observational scale σ. When the sample size n is large enough and $c \leq 1$, then (A.3) is summarized as the existence of negative moments of σ. In this case, the inverse-gamma distribution for σ^2, which is a typical choice of priors in many applications, satisfies (A.3).

2.5. Tail heaviness for robustness

Theorem 1 proves the posterior robustness of the linear regression models with the N-LPMN distributions, whose density tails are evaluated as $f_{\text{N-LPMN}}(x) \approx |x|^{-1}(\log|x|)^{-1-\gamma}$, as shown in Proposition 1. These super heavy tails are, in fact, necessary conditions for posterior robustness. To clarify the relationship between the posterior robustness and the tail behavior of the error distributions, we study a wider class of error distributions that includes the proposed distribution as a special case. This is defined by replacing $H(u;\gamma)$ in (3) with

$$H(u;\gamma,\delta) = C(\delta,\gamma)\frac{1}{(1+u)^{1+\delta}}\frac{1}{\{1+\log(1+u)\}^{1+\gamma}}, \quad u > 0, \tag{5}$$

where $C(\delta, \gamma)$ is a normalizing constant, and $\delta \geq 0$ is an additional shape parameter. Similar to the degree of freedom of t-distributions, the shape parameter δ is related to the decay of the density tail of (5); that is, $H(u; \gamma, b) \approx u^{-\delta-1}(\log u)^{-1-\gamma}$. Thus, this class of distributions covers the error distributions whose density tails are lighter than those of the proposed N-LPMN distribution in (3) and includes the N-LPMN distribution with the heaviest tails under $\delta = 0$. Note that the density tails become heavier than those of the Cauchy distribution if $\delta < 1$.

It is shown that the choice of a hyperparameter that can achieve posterior robustness is $\delta = 0$ (and arbitrary $\gamma > 0$). This is exactly the model considered in Theorem 1. From this observation, we conclude that the tails of the error distribution that are heavier than those of Cauchy distributions are essential for posterior robustness. For details, see the Supplementary Materials.

2.6. Existence of posterior moments

The N-LPMN distribution is too heavily tailed to have finite moments. However, the posterior of (β, σ^2) has finite means and variances in most situations. We verify this result for the inverse-gamma prior for σ^2.

Proposition 2. *Consider the linear regression model with the N-LPMN distribution in (3) and the prior for (β, σ) given in (4). Furthermore, suppose that the prior for σ^2 is an inverse-gamma distribution.*

(a) If (A.2) holds for some $c > 0$ and $c \leq n$, then $E[|\beta_k|^c|\mathcal{D}] < \infty$ for $k = 1, \ldots, p$.

(b) If $d \leq n$, then $E[\sigma^d|\mathcal{D}] < \infty$.

It is immediately apparent from (a) that the posterior means and variances of coefficients β exist under the horseshoe prior for β, which is given later in (6).

Corollary 1. *If the prior for β is horseshoe and $n \geq 2$, then $E[|\beta_k|^2|y] < \infty$.*

In fact, the existence of posterior moments of (β, σ^2) can be discussed for a broad class of error distributions and priors for (β, σ), not being limited to the linear regression model considered in this paper. Proposition 2 is proved with such a generality in the Supplementary Materials.

3. Posterior Computation

3.1. Gibbs sampler by augmentation

An important property of the proposed N-LPMN distribution (3) is its computational tractability, that is, we can easily construct a simple Gibbs sampler for posterior inference. Note that the error distribution contains two unknown parameters, s and γ. We can adopt conditionally conjugate priors given by $s \sim \text{Beta}(a_s, b_s)$ and $\gamma \sim \text{Ga}(a_\gamma, b_\gamma)$. The conditionally conjugate priors can also be found for main parameters, β and σ^2, and we use $\beta \sim N(A_\beta, B_\beta)$ and $\sigma^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$. The multivariate normal prior for β can be replaced with a scale mixture of normals, including shrinkage priors, which is discussed in Section 3.2.

To derive the tractable conditional posteriors, we need to keep the likelihood conditionally Gaussian with the latent variable u_i. This can be done easily by conditioning a set of latent variables (z_i, u_i). Consequently, the model is conditionally conjugate for our choice of priors for (β, σ^2) .

The full conditional distributions of the other parameters and latent variables in the LPMN distribution are not any standard distribution. However, we can augment the model with another set of latent variables by utilizing the following integral expression of density $H(u_i; \gamma)$:

$$H(u_i;\gamma) = \iint_{(0,\infty)^2} \text{Ga}(u_i; 1, v_i)\text{Ga}(v_i; w_i, 1)\text{Ga}(w_i; \gamma, 1)dv_i dw_i.$$

The random variable u_i, following the density $H(u_i; \gamma)$, admits the mixture representation: $u_i|(v_i, w_i) \sim \text{Ga}(1, v_i)$, $v_i|w_i \sim \text{Ga}(w_i, 1)$, and $w_i \sim \text{Ga}(\gamma, 1)$, which enables us to easily generate samples from the full conditional distribution of $(u_i|v_i, w_i)$ and $(v_i, w_i|u_i)$.

The introduction of the two latent states, (v_i, w_i), is useful in deriving the conditional posterior of u_i, and the algorithm of the Gibbs sampler immediately follows with latent (v_i, w_i) as the part of the Markov chain. However, (v_i, w_i) is redundant in the posterior sampling of the other parameters. We marginalize (v_i, w_i) out when sampling γ from its conditional posterior. This modification of the original Gibbs sampler simplifies the sampling procedure, and facilitates mixing, while targeting the same stationary distribution of the original Markov chain. The algorithm for posterior sampling is summarized as follows.

Summary of the posterior sampling

- Sample β from the full conditional distribution $N(\widetilde{B}\widetilde{A}, \widetilde{B})$, where

$$\widetilde{B}^{-1} = B_\beta^{-1} + \sigma^{-2} X^\top D X \quad \text{and} \quad \widetilde{A} = B_\beta^{-1} A_\beta + \sigma^{-2} X^\top D Y$$

  with $D = \mathrm{diag}(u_1^{-z_1}, \ldots, u_n^{-z_n})$.

- Sample σ^{-2} from $\mathrm{Ga}(\widetilde{a}_\sigma, \widetilde{b}_\sigma)$, where

$$\widetilde{a}_\sigma = a_\sigma + n/2 \quad \text{and} \quad \widetilde{b}_\sigma = b_\sigma + \sum_{i=1}^n (y_i - x_i^\top \beta)^2 / 2u_i^{z_i}.$$

- Sample z_i from the Bernoulli distribution; the probabilities of $z_i = 0$ and $z_i = 1$ are proportional to $(1-s)\phi(y_i; x_i^\top \beta, \sigma^2)$ and $s\phi(y_i; x_i^\top \beta, \sigma^2 u_i)$, respectively.

- The full conditional distribution of s is given by $\mathrm{Beta}(\widetilde{a}_s, \widetilde{b}_s)$, where $\widetilde{a}_s = a_s + \sum_{i=1}^n z_i$ and $\widetilde{b}_s = b_s + n - \sum_{i=1}^n z_i$.

- The full conditional distribution of $(\gamma, v_{1:n}, w_{1:n})$ is decomposed into those of γ and $(v_{1:n}, w_{1:n}|\gamma)$.

  - The full conditional distribution of γ (with $v_{1:n}$ and $w_{1:n}$ marginalized out) is given by $\mathrm{Ga}(\widetilde{a}_\gamma, \widetilde{b}_\gamma)$, where $\widetilde{a}_\gamma = a_\gamma + n$ and $\widetilde{b}_\gamma = b_\gamma + \sum_{i=1}^n \log\{1 + \log(1 + u_i)\}$.
  - The full conditional distributions of $(v_1, w_1), \ldots, (v_n, w_n)$ are mutually independent. For each i, (v_i, w_i) can be sampled in a compositional manner. Sample w_i from $\mathrm{Ga}(1+\gamma, 1+\log(1+u_i))$, then sample $(v_i|w_i)$ as $\mathrm{Ga}(1 + w_i, 1 + u_i)$.

- The full conditional distribution of u_i is $\mathrm{GIG}(1/2, 2v_i, (y_i - x_i^\top \beta)^2/\sigma^2)$ if $z_i = 1$, or $\mathrm{Ga}(1, v_i)$ if $z_i = 0$.

Finally, we remark on the choice of hyperparameters in the priors for s and γ. Although the LPMN distribution is log-regularly varying under arbitrary $\gamma > 0$, the use of a large value of γ is not suitable for capturing potential outliers. This is because the tail of the LPMN distribution becomes lighter as γ increases. Moreover, the use of different values of γ would not

considerably affect the posterior result, given that γ is not large. Hence, instead of using a diffuse prior for γ, we recommend simply using a fixed value. We adopt $\gamma = 1$ as the default choice, and the sensitivity of this choice is investigated in Section 4. As a more data-dependent method, we also recommend employing an informative prior that prevents large values of γ by setting, for example, $a_\gamma = b_\gamma = 100$, which is considered in Section 4. Regarding the mixing proportion s, we adopt $a_s = b_s = 1$, resulting in a uniform prior for s as the default choice.

3.2. Robust Bayesian variable selection with shrinkage priors

When the dimension of x_i is moderate or large, it is desirable to select a suitable subset of x_i to achieve an efficient estimation. This procedure of variable selection is also seriously affected by possible outliers, by which we may fail to select suitable subsets of covariates. For a robust Bayesian variable selection procedure, we introduce shrinkage priors for the regression coefficients. Here, we rewrite the regression model to explicitly express an intercept term as $y_i = \alpha + x_i^t\beta + \varepsilon_i$, and consider a normal prior $\alpha \sim N(0, A_\alpha)$ with a fixed hyperparameter $A_\alpha > 0$. For the regression coefficients β, we consider a class of independent priors expressed as a scale mixture of normals, given by:

$$\pi(\beta) = \prod_{k=1}^{p} \int_0^\infty \phi(\beta_k; 0, \sigma^2\tau^2\xi_k)\pi_\xi(\xi_k)\mathrm{d}\xi_k, \tag{6}$$

where $\pi_\xi(\cdot)$ is a mixing distribution, and τ^2 is an unknown global parameter that controls the strength of the shrinkage effects. Examples of the mixing distribution $\pi_\xi(\cdot)$ include the exponential distribution leading to the Laplace prior of β (Bayesian lasso, Park and Casella 2008), and the half-Cauchy distribution for $\xi_k^{1/2}$ which results in the horseshoe prior (Carvalho et al., 2009, 2010). The robustness property of the resulting posterior distributions is guaranteed for those shrinkage priors because Assumption (A.2) of Theorem 1 is satisfied.

In terms of posterior computation, the key property is that the conditional distribution of β_k given ξ_k under (6) is a normal distribution, so the sampling algorithm given in Section 3.1 is still valid with minor modifications. Specifically, the sampling from the full conditional distributions of α, β, σ^2, and $\xi_1, \ldots, \xi_p$ is modified or newly added as follows:

- Sample α from $N(\widetilde{A}_\alpha^{-1}\widetilde{B}_\alpha, \widetilde{A}_\alpha^{-1})$, where

$$\widetilde{A}_\alpha = A_\alpha + \sigma^{-2}\sum_{i=1}^{n} u_i^{-1} \quad \text{and} \quad \widetilde{B}_\alpha = \sigma^{-2}\sum_{i=1}^{n} u_i^{-1}(y_i - x_i^\top \beta).$$

- Sample β from $N(\widetilde{A}_\beta^{-1}X^\top D\widetilde{Y}, \sigma^2\widetilde{A}_\beta^{-1})$, where

$$\widetilde{Y} = Y - \alpha 1_n \quad \text{and} \quad \widetilde{A}_\beta = \Lambda^{-1} + X^\top DX \quad \text{with} \quad \Lambda = \tau^2 \text{diag}(\xi_1, \ldots, \xi_p).$$

- Sample σ^{-2} from $\text{Ga}(\widetilde{a}_\sigma, \widetilde{b}_\sigma)$, where

$$\widetilde{a}_\sigma = a_\sigma + (n+p)/2 \quad \text{and} \quad \widetilde{b}_\sigma = b_\sigma + \sum_{i=1}^{n}(y_i - x_i^\top \beta)^2/2u_i^{z_i} + \beta^\top \Lambda^{-1}\beta.$$

- Sample ξ_k and τ^2 from their full conditionals. Their densities are proportional to $\phi(\beta_k; 0, \sigma^2\tau^2\xi_k)\pi_\xi(\xi_k)$ and $\pi_{\tau^2}(\tau^2)\prod_{k=1}^{p}\phi(\beta_k; 0, \sigma^2\tau^2\xi_k)$, respectively, where $\pi_{\tau^2}(\tau^2)$ is the prior density for τ^2.

The normal mixture representation of the N-LPMN distribution and shrinkage priors makes the full conditional distributions of α and β computationally tractable. The sampling of ξ_k and τ^2 depends on the choice of shrinkage priors, but the existing algorithms in the literature can be directly imported to our method.

In Section 5, we adopt the horseshoe prior for the regression coefficients with the N-LPMN distribution for the error terms. Here, we provide details of the sampling algorithm under the horseshoe model. The horseshoe prior assumes that $\sqrt{\xi_k} \sim C^+(0,1)$ independently for $k = 1, \ldots, p$ and $\tau \sim C^+(0,1)$, where $C^+(0,1)$ is the standard half-Cauchy distribution with probability density function given by $p(x) = 2/\pi(1+x^2)$ for $x > 0$. Note that they admit the hierarchical expressions given by $\xi_k|\lambda_k \sim \text{IG}(1/2, 1/\lambda_k)$ and $\lambda_k \sim \text{IG}(1/2, 1/2)$ for ξ_k, and $\tau^2|\nu \sim \text{IG}(1/2, 1/\nu)$ and $\nu \sim \text{IG}(1/2, 1/2)$ for τ^2. Then, we can sample from each full conditional distribution as follows:

- Sample ξ_k from $\text{IG}(1, 1/\lambda_k + \beta_k^2/2\tau^2\sigma^2)$.

- Sample λ_k from $\text{IG}(1, 1 + 1/\xi_k)$.

- Sample τ^2 from $\text{IG}((p+1)/2, 1/\nu + \sum_{k=1}^{p}\beta_k^2/2\xi_k\sigma^2)$.

- Sample ν from $\mathrm{IG}(1, 1 + 1/\tau^2)$.

These sampling steps can be directly incorporated into the Gibbs sampling algorithm described in Section 3.1.

3.3. Hierarchical linear regression

The proposed error distribution can be adopted in more general linear regression models. As an example, we consider a hierarchical model given by

$$y_i = x_i^\top \beta + g_i^\top b + \varepsilon_i, \quad i = 1, \ldots, n, \tag{7}$$

where g_i is an r-vector of additional covariates and b is a vector of random effects distributed as $b \sim N(0, H(\psi))$ with $r \times r$ covariance matrix $H(\psi)$ parametrized by ψ. To absorb the potential effects of outliers, we use the N-LPMN distribution for ε_i. The model structure described in (7) is general enough to represent a wide variety of useful models, as discussed in later sections. Even under model (7), the robustness properties for β, as discussed in Section 2.4, can be proven by checking whether the prior for b satisfies Assumption (A.2). Moreover, the augmentation strategy for the efficient posterior computation algorithm can still be employed, and the full conditional distribution of b is normal. We adopt a random intercept model for longitudinal data in our simulation study in Section 4.3 and a linear regression with spatial effects in our application in Section 5.1.

4. Simulation studies

4.1. Linear regression with small n and p

In this study, we carry out simulation studies to investigate the performance of the proposed method together with existing methods. We first consider $n = 50$ observations from the linear regression model with $p = 3$ covariates, given by

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \sigma \varepsilon_i, \quad i = 1, \ldots, n, \tag{8}$$

where $\beta_0 = 0.5, \beta_1 = 0.3, \beta_2 = 0, \beta_3 = 0.3$ and $\sigma = 0.5$. Here the vector of covariates $(x_{i1}, \ldots, x_{ip})$ is generated from a multivariate normal distribution with zero mean vector and variance-covariance matrix whose (k, ℓ)-entry has

$(0.2)^{|k-\ell|}$ for $k, \ell \in \{1, \ldots, p\}$. Regarding the contamination structure of the error term, we adopt the location-shift model (Abraham and Box, 1978):

$$\varepsilon_i \sim (1-\omega)N(0,1) + \omega N(\mu, 1), \quad i = 1, \ldots, n,$$

where ω is the contamination ratio and μ is the location of the outliers. We consider all combinations of $\omega \in \{0.05, 0.1, 0.2\}$ and $\mu \in \{10, 20\}$, in addition to the case of no contamination ($\omega = 0$), which leads to seven scenarios in total. In this setting, we replicate 20000 datasets independently.

The error distributions we consider include the N-LPMN distribution, the LPTN distribution (Gagnon et al., 2019), and t-distribution with ν degrees of freedom. For the hyperparameter γ in the N-LPMN distribution, we fix $\gamma = 1$ (denoted by N-LP in the tables) and estimated γ adaptively (aN-LP) by assigning a $\mathrm{Ga}(100, 100)$ prior distribution. In both N-LPMN models, the mixture weight s follows the uniform distribution on $[0, 1]$, as explained in Section 3.1. For the LPTN distribution, the tuning parameter $\rho \in (2\Phi(1) - 1, 1) \approx (0.6827, 1)$ is specified as $\rho = 0.95$ and $\rho = 0.8$ (LP1 and LP2, respectively). Regarding the degree of freedom ν in the t-distribution, we select the results of $\nu = 1$ (Cauchy distribution, denoted by C) and $\nu = 3$ (T3). Similarly, we consider an adaptive version (aT) that employs a discrete uniform prior on $\nu \in \{1, 2, 3, 4, 5, 8, 10, 15, 20, 30, 50\}$. In addition, the two-component mixture of the standard normal distribution and the t-distribution with $\nu = 1/2$ is considered (MT), with the uniform prior for the mixture weight as in the N-LPMN models. For comparison, we also adopt the normal distribution as the error distribution (denoted by N), which should perform best in the absence of outliers. Note that all the error distributions listed here are "misspecified" because they do not include the location shift of the error term in the data generating process. This setting emphasizes that the posterior robustness verified in this research is valid regardless of the structure of the outliers.

The priors for the regression coefficients and observational scale are set as $\beta_k \sim N(0, 1000)$ and $\sigma^{-2} \sim \mathrm{Ga}(0.1, 0.1)$ for all the models. To employ the posterior inference, we generated the posterior samples of (β, σ) using a Gibbs sampler under the N-LPMN, t and normal error distributions. For the LPTN distribution, the random-walk Metropolis-Hastings algorithm is adopted, as in Gagnon et al. (2019), in which the step sizes are set to 0.05. For each of the nine models, we generate 1000 posterior samples after discarding the first 500 samples.

Based on the posterior samples, we compute the posterior means and 95% credible intervals of β_k for $k = 1, \ldots, p$. The performance of the point and interval estimation is assessed by the square root of the mean squared errors (RMSE), coverage probabilities (CP), and average length (AL) as based on 20000 replications of the simulation. These values are then averaged over $\beta_0, \ldots, \beta_p$. We also evaluate the RMSE of σ. To measure the efficiency of the sampling schemes, we compute the average of the inefficiency factors (IF) of the posterior samples, defined as, $1 + 2\sum_{k=1}^{\infty} \rho_k$, where ρ_k is the lag-k autocorrelation of the posterior samples. We used the `numEff` function available in the R package "bayesm".

In Table 1, we report the values of these performance measures for seven scenarios. When $\omega = 0$ (no outlier), the normal error distribution provides the smallest RMSE and IF. While the other methods are slightly inefficient, the proposed method (N-LP and aN-LP in the table) performs almost in the same way as the normal distribution. This is empirical evidence that the efficiency loss of the N-LPMN distribution is negligible owing to the structure of the two-component mixture. In the other robust methods, the RMSEs are slightly higher than that of the normal distribution, and the CPs are smaller than the nominal level.

In the other scenarios, where outliers are incorporated in the data generating process, the performance of the normal distribution is significantly lowered, and the robustness property of the other models becomes evident in the performance measures. In particular, the N-LPMN distribution with a fixed γ (N-LP) performs quite stably in both point and interval estimations. The adaptive version (aN-LP) also works reasonably well, and the performance is comparable to that of the N-LPMN. The LPTN model with $\rho = 0.95$ (LP1) shows reasonable performance in point estimation, but its CPs tend to be smaller than the nominal level. The other LPTN model with $\rho = 0.8$ (LP2) worsens the accuracy of point estimation, implying the sensitivity of the choice of hyperparameter ρ to the posteriors. The T3 and aT models also suffer from larger RMSE values, especially in the scenarios of large ω and μ, which emphasizes the lack of posterior robustness under the t-distribution family. In addition, the interval estimation under the t-distributions depends on the degree-of-freedom parameter. This is also seen in the results of the Cauchy and t_3-distributions, where the credible intervals are too wide and narrow, respectively. The MT model is competitive with the N-LPMN and LP1 models in most scenarios, but its performance measures significantly deteriorate when the contamination level is high ($\omega = 0.2$).

Interestingly, the Cauchy model (C) provides the smallest RMSE in both β and σ when $\omega = 0.2$. We give discussions on this issue in the Supplementary Material.

In terms of computational efficiency, the IF values of the N-LPMN models are small and comparable with those of the t-distribution methods, as expected from its simple Gibbs sampling algorithm. Meanwhile, the IFs of the LPTN models are very large because of the use of the Metropolis-Hastings algorithm. This result empirically shows that for a reliable posterior analysis under the LPTN models, the number of iterations in the computation by MCMC must be increased. In this context, more effort is needed to tune the step-size parameter. Furthermore, we measure the raw computation time of five methods (N-LP, LP1, T3, MT and N) for several sample sizes n, which are reported in the Supplementary Materials. The robust models (N-LP and LP1) require a longer computational time than the simpler models (T3, MT, and N) due to the complexity of their models. Hence, the lower IFs of the N-LPMN model confirmed in this simulation study are important to complement the longer computational time per iteration.

4.2. Moderately large n and p

We consider a setting with a larger sample size and many predictors, $n = 300$ and $p = 20$. In doing so, we employ model (8), where $\beta_0 = 0.5, \beta_1 = \beta_4 = 0.3$, $\beta_7 = \beta_{10} = 2$, $\sigma = 0.5$ and the other coefficients are set to 0. We consider all combinations of $\omega \in \{0.05, 0.1, 0.2\}$ and $\mu \in \{10, 20\}$, in addition to the case of no contamination ($\omega = 0$), which leads to 7 scenarios in total. In this setting, we replicate 1000 independent datasets.

In Table 2, we report the values of these performance measures for nine scenarios. The increased sample size highlights the robustness of the N-LPMN model more clearly. The RMSE (β) of N-LPMN is smaller than that of C, and the RMSE (β) of N-LPMN is smaller than that of MT in all scenarios. The coefficient vector β is now $p = 20$ dimensional, which becomes the computational burden for the LP1 and LP2 models that use the Metropolis-Hastings algorithm in posterior sampling, as seen in the higher values of IFs.

Finally, we evaluate the predictive performance. We generate $m = 20$ additional covariates x_{j*} $(j = 1, \ldots, m)$ from the same multivariate normal distribution, and then generated the true response value y_{j*} based on the linear regression with $\varepsilon_i \sim N(0, 1)$. In other words, the predicted response is not contaminated with outliers. Accordingly, in the prediction using the

Table 1: Average values of RMSEs, CPs, ALs and IFs of the proposed N-LPMN distribution with γ fixed (N-LP) and estimated (aN-LP), log-Pareto-tailed normal distribution with $\rho = 0.95$ (LP1) and $\rho = 0.8$ (LP2), Cauchy distribution (C), t-distribution with three degrees of freedom (T3) and estimated degrees of freedom (aT), two-component mixture of normal and t-distribution with 1/2 degrees of freedom (MT), and normal linear regression (N), based on 20000 replications in seven combinations of $(100\omega, \mu)$ with $n = 50$ and $p = 3$. RMSE and AL are multiplied by 10. The best RMSE values are highlighted in bold.

	$(100\omega, \mu)$	N-LP	aN-LP	LP1	LP2	C	T3	aT	MT	N
RMSE (β)	(0, –)	**0.76**	**0.76**	0.77	0.80	0.91	0.81	0.78	**0.76**	**0.76**
	(5, 10)	**0.80**	0.81	0.83	0.82	0.93	0.84	0.92	**0.80**	2.20
	(10, 10)	**0.87**	**0.87**	1.70	0.86	0.94	0.97	1.42	1.11	3.47
	(20, 10)	1.72	1.81	5.62	2.07	**1.10**	2.72	3.67	4.53	5.90
	(5, 20)	**0.80**	**0.80**	**0.80**	0.82	0.92	0.82	0.93	**0.80**	4.23
	(10, 20)	0.86	0.86	0.91	**0.84**	0.93	0.91	1.79	1.50	6.80
	(20, 20)	1.26	1.28	5.30	**0.98**	1.02	4.34	6.48	8.71	11.7
RMSE (σ)	(0, –)	**0.53**	**0.53**	0.58	1.04	1.92	1.05	0.81	**0.53**	**0.53**
	(5, 10)	**0.57**	0.58	1.06	1.20	1.67	0.81	1.69	0.62	7.48
	(10, 10)	**0.73**	0.75	4.29	1.45	1.37	1.85	4.11	2.03	11.2
	(20, 10)	3.44	3.77	14.5	4.93	**1.16**	7.19	9.96	11.3	16.0
	(5, 20)	**0.57**	**0.57**	0.77	1.15	1.67	0.86	2.84	0.65	18.0
	(10, 20)	**0.63**	**0.64**	1.69	1.31	1.37	2.40	7.67	3.83	25.9
	(20, 20)	2.07	2.22	15.8	1.98	**1.21**	14.3	21.5	24.4	35.9
CP (%)	(0, –)	95.0	95.0	93.9	92.7	89.7	93.5	94.7	95.0	95.1
	(5, 10)	94.7	94.6	94.9	93.3	91.2	95.7	97.5	94.8	91.3
	(10, 10)	94.3	94.2	93.8	93.8	92.8	97.2	98.0	94.5	81.7
	(20, 10)	93.3	93.3	74.8	93.7	95.3	92.5	90.3	85.2	71.8
	(5, 20)	94.7	94.6	94.7	93.1	91.3	96.0	98.3	94.9	91.0
	(10, 20)	94.2	94.1	95.7	93.6	92.9	97.8	99.3	94.6	80.3
	(20, 20)	93.9	93.8	91.2	94.7	95.9	95.7	94.2	86.3	71.7
AL	(0, –)	3.02	3.02	3.00	2.98	3.03	3.03	3.07	3.02	3.02
	(5, 10)	3.18	3.18	3.34	3.15	3.22	3.48	4.35	3.17	6.99
	(10, 10)	3.37	3.38	4.29	3.37	3.48	4.30	6.36	3.52	9.38
	(20, 10)	4.25	4.33	8.98	4.39	4.31	8.18	11.0	8.18	12.3
	(5, 20)	3.16	3.16	3.24	3.13	3.22	3.49	5.12	3.17	12.8
	(10, 20)	3.33	3.33	3.62	3.30	3.47	4.39	9.30	3.68	17.9
	(20, 20)	3.83	3.84	6.91	3.80	4.24	11.8	20.3	13.6	24.2
IF	(0, –)	1.20	1.22	16.1	17.0	4.32	2.11	1.85	1.07	1.02
	(5, 10)	2.24	2.37	17.4	17.7	3.99	1.92	1.86	1.43	1.02
	(10, 10)	3.33	3.51	20.4	18.5	3.68	1.82	2.11	2.00	1.02
	(20, 10)	4.87	4.95	31.5	21.6	3.17	2.17	2.44	3.13	1.02
	(5, 20)	2.25	2.36	17.1	17.6	3.97	1.89	1.91	1.42	1.02
	(10, 20)	3.35	3.50	18.5	18.2	3.65	1.73	2.41	1.99	1.02
	(20, 20)	4.86	4.89	25.3	20.0	3.00	2.25	2.99	3.27	1.02

N-LPMN and MT distributions, we construct the sampling model of y_{j*} conditional on $z_j = 0$ as

$$f(y_{j*}|\mathcal{D}, z_j = 0) = \int \phi(y_{j*}; x_{j*}^\top \beta, \sigma^2)\pi(\beta, \sigma|\mathcal{D})d\beta d\sigma.$$

This predictive distribution reflects our belief that prediction should be considered only for non-outlying observations. If one believes that the predicted response might also be outlying, then the model in (3) can be used for prediction without conditioning z_j at the cost of inflated predictive uncertainty. To handle the outlying predictive values, however, the models for outlier detection should be more appropriate. For the LPTN and t-distributions, it is difficult to separate the non-outliers and outliers. For these models, we use the same error distributions for prediction. We report the result of the T3 model only; the 95% predictive intervals of y_{j*} under the LPTN and other t models are extremely wide due to their (super) heavy tails.

To evaluate the predictive performance, we compute the MSE of the posterior predictive mean and CP and AL of 95% predictive intervals of y_{j*}. These values are averaged over 1000 replications, as shown in Table 3. It can be seen that the model with the Gaussian errors produces worse point predictions and wider interval estimates as more and larger outliers are generated, which is clearly due to the lack of posterior robustness. The other robust methods are equally performative in terms of point prediction, but they show a significant difference in uncertainty quantification. The T3 method tends to be too conservative, in the sense that the predictive intervals are too wide and show almost 100% coverage. The N-LPMN and MT models have similar predictive results, whereas the coverage rates suggest the potential under-coverage of the MT model. This result shows the importance of posterior robustness and the use of error distributions with super heavy tails in estimation for both posterior inference and predictive analysis.

4.3. Random intercept models

Next, we consider simulation studies using the following random intercept model:

$$y_{jt} = \beta_0 + \sum_{k=1}^{p} \beta_k x_{jtk} + v_j + \sigma\varepsilon_{jt}, \quad t = 1, \ldots, T, \quad j = 1, \ldots, m, \qquad (9)$$

Table 2: Average values of RMSEs, CPs, ALs and IFs under larger scale simulation studies ($n = 300$ and $p = 20$). The methods to be compared are the proposed N-LPMN distribution with γ fixed (N-LP) and estimated (aN-LP), log-Pareto-tailed normal distribution with $\rho = 0.95$ (LP1) and $\rho = 0.8$ (LP2), Cauchy distribution (C), t-distribution with three degrees of freedom (T3) and estimated degrees of freedom (aT), two-component mixture of normal and t-distribution with 1/2 degrees of freedom (MT), and normal linear regression (N), based on 1000 replications in seven combinations of $(100\omega, \mu)$. The RMSE and AL are multiplied by 10. The best RMSE values are highlighted in bold.

	$(100\omega, \mu)$	N-LP	aN-LP	LP1	LP2	C	T3	aT	MT	N
RMSE (β)	(0, –)	**0.31**	**0.31**	0.33	0.34	0.38	0.33	0.32	**0.31**	**0.31**
	(5, 10)	0.33	0.33	0.34	0.35	0.38	0.34	0.36	**0.32**	0.92
	(10, 10)	**0.35**	0.36	0.37	0.37	0.39	0.37	0.50	0.41	1.47
	(20, 10)	**0.41**	0.43	2.05	**0.41**	0.42	0.78	1.38	2.51	2.53
	(5, 20)	0.33	0.33	0.34	0.35	0.38	0.34	0.34	**0.32**	1.78
	(10, 20)	**0.35**	**0.35**	0.36	0.37	0.39	**0.35**	0.52	0.42	2.88
	(20, 20)	**0.40**	0.41	0.43	**0.40**	**0.40**	0.75	2.23	5.00	5.04
RMSE (σ)	(0, –)	**0.20**	0.21	0.23	0.74	1.82	0.97	0.70	**0.20**	0.21
	(5, 10)	**0.22**	**0.22**	0.45	0.87	1.57	0.41	0.93	0.26	7.02
	(10, 10)	**0.26**	0.34	0.93	1.06	1.26	0.85	3.10	1.50	10.9
	(20, 10)	0.79	1.94	14.0	1.62	**0.47**	5.60	9.24	15.5	15.6
	(5, 20)	**0.22**	**0.22**	0.38	0.83	1.56	0.40	1.45	0.26	17.4
	(10, 20)	**0.25**	0.27	0.67	0.99	1.25	0.93	5.34	2.01	25.4
	(20, 20)	**0.48**	0.77	2.69	1.33	**0.48**	8.60	19.4	35.0	35.3
CP (%)	(0, –)	95.2	95.2	90.2	88.0	89.9	93.8	94.7	95.0	95.2
	(5, 10)	94.7	94.8	90.9	87.9	91.4	95.7	97.7	94.8	90.1
	(10, 10)	94.3	94.1	91.7	88.6	92.6	97.2	98.2	94.7	90.6
	(20, 10)	93.8	93.0	75.0	89.4	95.3	95.1	94.4	90.3	90.3
	(5, 20)	94.8	94.7	90.6	88.2	91.3	95.9	98.7	95.0	90.2
	(10, 20)	94.8	94.6	91.8	88.8	93.3	98.2	99.6	95.2	90.3
	(20, 20)	93.6	93.2	93.0	89.3	96.1	97.8	96.4	90.1	90.0
AL	(0, –)	1.22	1.22	1.13	1.11	1.25	1.24	1.25	1.22	1.23
	(5, 10)	1.28	1.28	1.22	1.16	1.33	1.40	1.69	1.28	2.89
	(10, 10)	1.35	1.36	1.36	1.23	1.41	1.64	2.46	1.38	3.84
	(20, 10)	1.55	1.59	2.78	1.40	1.67	2.93	4.40	4.96	5.00
	(5, 20)	1.28	1.28	1.21	1.16	1.32	1.40	1.86	1.28	5.40
	(10, 20)	1.34	1.34	1.29	1.21	1.41	1.65	3.44	1.37	7.36
	(20, 20)	1.50	1.51	1.62	1.35	1.66	3.46	8.03	9.68	9.77
IF	(0, –)	1.00	1.00	31.1	31.9	4.10	2.02	1.78	0.97	0.97
	(5, 10)	1.87	2.18	31.7	32.2	3.80	1.83	1.72	1.24	0.97
	(10, 10)	2.92	3.58	32.6	32.5	3.47	1.65	1.88	1.55	0.97
	(20, 10)	5.27	6.05	38.1	33.5	2.87	1.61	2.07	0.99	0.97
	(5, 20)	1.88	2.14	31.6	32.2	3.78	1.81	1.70	1.25	0.97
	(10, 20)	2.91	3.33	32.2	32.5	3.45	1.60	2.07	1.52	0.97
	(20, 20)	5.36	5.86	34.2	33.2	2.80	1.42	2.54	1.04	0.97

Table 3: Average values of RMSEs of posterior predictive means and CPs and ALs of 95% prediction intervals based on the N-LPMN method with $\gamma = 1$ (N-LP), t-distribution with three degrees of freedom (T3), two-component mixture of normal and t-distribution with 1/2 degrees of freedom (MT), and the standard normal linear regression (N), based on 1000 replications in seven combinations of $(100\omega, \mu)$. The RMSE and CP are multiplied by 100. The best RMSE values are highlighted in bold.

	$(100\omega, \mu)$	N-LP	T3	MT	N
RMSE	(0, –)	**51.7**	52.1	**51.7**	**51.7**
	(5, 10)	**52.2**	52.5	**52.2**	65.1
	(10, 10)	**52.4**	52.7	53.3	82.6
	(20, 10)	**53.5**	61.6	125.8	126.8
	(5, 20)	**52.2**	52.4	**52.2**	93.7
	(10, 20)	**52.2**	52.3	54.4	139.9
	(20, 20)	**52.9**	61.1	229.9	233.2
CP	(0, –)	94.9	98.5	94.9	95.1
	(5, 10)	94.9	99.5	94.3	100.0
	(10, 10)	95.0	99.9	93.7	100.0
	(20, 10)	96.8	100.0	99.9	100.0
	(5, 20)	94.6	99.4	94.2	100.0
	(10, 20)	94.9	99.9	93.9	100.0
	(20, 20)	96.1	100.0	99.9	100.0
AL	(0, –)	2.0	2.6	2.0	2.1
	(5, 10)	2.0	3.0	2.0	4.8
	(10, 10)	2.1	3.7	2.0	6.4
	(20, 10)	2.3	6.6	8.3	8.4
	(5, 20)	2.0	3.0	2.0	9.0
	(10, 20)	2.1	3.7	2.1	12.3
	(20, 20)	2.2	8.1	16.0	16.3

where $v_j \sim N(0, \tau_v^2)$ is a random effect. This is an example of the general model presented in Section 3.3. The model of this type is frequently used in longitudinal data analysis (e.g. Verbeke, 2009), where m and T are the numbers of subjects and repeated measurements, respectively, and v_j is regarded as a subject-specific effect. Throughout this study, we set $m = 50$, $T = 10$ and $p = 10$. We adopt the same values for β_k, and the same generating process for $(x_{jt1}, \ldots, x_{jtp})$ and ε_{jt}, as those in the previous simulation study. The other parameters are set as $\tau_v^2 = (0.5)^2$ and $\sigma = 1$.

We model the distribution of error ϵ_{jt} in the model (9) by the N-LPMN distribution with latent variables (z_{jt}, u_{jt}). The same data augmentation strategy can be used in the posterior computation for this model, and the full conditional distribution of v_j is given by $N(\widetilde{b}_j\widetilde{a}_j, \widetilde{b}_j)$, where

$$\widetilde{b}_j^{-1} = \frac{1}{\tau_v^2} + \frac{1}{\sigma^2}\sum_{t=1}^{T}\frac{1}{u_{jt}^{z_{jt}}} \quad \text{and} \quad \widetilde{a}_j = \frac{1}{\sigma^2}\sum_{t=1}^{T} u_{jt}^{-z_{jt}}\Big(y_{tj} - \beta_0 - \sum_{k=1}^{p}\beta_k x_{jtk}\Big).$$

We use an inverse-gamma prior for τ_v^2, namely, $\tau_v^2 \sim \mathrm{IG}(a_v, b_v)$ with $a_v = b_v = 1$, and the full conditional distribution of τ_v^2 is $\mathrm{IG}(\widetilde{a}_v, \widetilde{b}_v)$, where $\widetilde{a}_v = a_v + m/2$ and $\widetilde{b}_v = b_v + \sum_{j=1}^{m} v_j^2/2$. Given the random effect v_j, other parameters and latent variables can be easily generated from their full conditional distributions in Section 3.1, with a slight modification by replacing the response variable with $y_{jt} - v_j$. The other error distributions, such as the normal and t-distributions and the finite mixture, can be implemented in the same way by using its representation of scale mixture of normals. The only exception is the LPTN distribution; it does not admit the representation of a scale mixture of normals and is not directly incorporated into the random intercept model. In total, we employ six error distributions (N-LPMN, aN-LPMN, C, aT, MT, and N) in this study. We evaluate the performance of point and interval estimations by posterior means and 95% credible intervals for the regression coefficients, using RMSE, CP, and AL, as adopted in the previous study. The performance of the six models in predicting the random effect is also assessed via the square root of the mean squared prediction error (RMSPE) based on 500 replications of the simulations, and these values are averaged over $v_1, \ldots, v_m$.

The results are listed in Table 4. Regarding the regression coefficients, a similar tendency as found in Tables 2 can be observed. This indicates the usefulness of the proposed N-LPMN method under more structured models

than linear regression. It is also observed that the N-LPMN method with estimated γ does not necessarily work well; therefore, our recommendation in this example is simply to use the fixed value $\gamma = 1$. In terms of the RMSPE, the proposed N-LPMN method consistently outperforms the other methods. Specifically, the difference between the N-LPMN and MT methods is considerable, which also suggests the importance of the posterior robustness shown in Theorem 1, that is, the advantage of the proposed error distribution over the conventional finite mixture approach with t-distribution.

5. Real data examples

The posterior robustness of the proposed N-LPMN distribution is demonstrated via the analysis of two real datasets: the Boston housing data and diabetes data. The goal of statistical analysis here is variable selection with $p = 29$ and $p = 64$ predictors in the presence or absence of outliers. Our robustness scheme is a prominent part of such analysis as it allows the use of unbounded prior densities for strong shrinkage effect. Specifically, this relates to the horseshoe priors discussed in Section 3.2, while protecting the posteriors from the potential outliers. The Boston housing data are suspected to be contaminated with outliers, where the difference of the proposed N-LPMN distribution and the traditional t-distribution is emphasized. In contrast, the diabetes data are free from extreme outliers, and we use this dataset to discuss the possible efficiency loss caused by the use by using N-LPMN distributions.

In our examples, the number of covariates is not small. Hence, we consider both variable selection and robust Bayesian inference using the proposed method. Specifically, we employed the horseshoe prior, as described in Section 3.2. For comparison, we also apply the standard normal distribution and the two-component mixture of normal and t-distributions as the error distribution, while using the horseshoe prior for regression coefficients. In all the methods, we generate 10000 posterior samples after discarding the first 5000 posterior samples as burn-in.

5.1. Boston housing data

We first consider the famous Boston housing dataset (Harrison and Rubinfeld, 1978). The response variable is the corrected median value of owner-occupied homes (1,000 USD). The covariates in the original datasets consist of 14 continuous-valued variables regarding the information of houses, such

Table 4: Average values of RMSEs, CPs, ALs and RMSPEs of the proposed N-LPMN distribution with γ fixed (N-LP) and estimated (aN-LP), Cauchy distribution (C), t-distribution with estimated degrees of freedom (aT), two-component mixture of normal and t-distribution with 1/2 degrees of freedom (MT), and normal distribution (N) under the random intercept models with six combinations of $(100\omega, \mu)$. All values are multiplied by 100. The best RMSE and RMSPE values are in bold.

	$(100\omega, \mu)$	N-LP	aN-LP	C	aT	MT	N
RMSE (fixed effects)	(5, 5)	5.91	**5.89**	6.86	6.54	5.91	10.67
	(10, 5)	8.45	8.88	**7.12**	9.39	8.52	17.51
	(5, 10)	5.61	**5.58**	6.84	6.37	5.72	19.40
	(10, 10)	5.86	**5.79**	6.78	9.52	6.03	33.74
	(5, 15)	5.47	**5.45**	6.65	6.10	5.58	28.23
	(10, 15)	5.86	**5.79**	6.84	9.36	5.96	49.80
CP	(5, 5)	94.1	94.0	81.6	92.6	92.1	86.5
	(10, 5)	92.9	93.2	84.5	90.4	91.1	85.7
	(5, 10)	95.1	94.7	82.3	95.2	91.9	85.9
	(10, 10)	95.2	95.2	86.0	95.3	91.8	86.0
	(5, 15)	94.9	94.6	83.6	96.4	91.9	86.4
	(10, 15)	95.5	95.4	84.8	97.3	92.5	86.5
AL	(5, 5)	22.0	21.8	18.4	22.7	20.3	27.7
	(10, 5)	25.8	26.0	19.7	28.5	23.1	33.7
	(5, 10)	21.4	21.1	18.5	25.7	19.7	44.7
	(10, 10)	23.1	22.7	19.6	38.9	21.0	58.5
	(5, 15)	21.3	21.0	18.5	27.2	19.7	63.3
	(10, 15)	23.0	22.6	19.6	47.2	20.9	85.1
RMSPE (random effects)	(5, 5)	29.5	**29.4**	33.9	31.5	35.0	40.5
	(10, 5)	33.6	**33.2**	34.1	37.1	38.8	44.2
	(5, 10)	**28.8**	**28.8**	34.2	33.3	33.7	46.9
	(10, 10)	30.5	**29.7**	33.5	40.9	36.4	48.3
	(5, 15)	**28.8**	**28.8**	34.0	33.9	33.6	48.3
	(10, 15)	30.3	**29.5**	33.4	41.8	36.3	49.2

as per capita crime rate and accessibility to radial highways, and one binary covariate. After standardizing the 14 continuous covariates, we use them to create squared values, which results in $p = 29$ covariates in our models. The sample size is 506. The data also contain the longitude and latitude of house i, denoted by t_i. To consider spatial correlation, we consider the following model:

$$y_i = x_i^\top \beta + g(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{10}$$

where $g(t_i)$ is a spatial effect as an unknown function of location information t_i. We assume that $g(t_i)$ follows the standard Gaussian process, namely, $\eta \equiv (g(t_1), \ldots, g(t_n))$ and $\eta \sim N(0, \kappa^2 C(h))$, where $C(h)$ is a variance-covariance matrix whose (i, j)-entry is $\exp(-\|s_i - s_j\|^2/2h^2)$ with unknown bandwidth parameter h. The above model can be seen as a spatially varying intercept model, or a spatially varying coefficient model (e.g. Gelfand et al., 2003). This is another example of the general model in Section 3.3, where $r = n$, $b = \eta$, g_i is the i-th standard basis, and $H(\psi) = \kappa^2 C(h)$ with $\psi = (\kappa, h)$. Under the N-LPMN distribution for ε_i, the full conditional distribution of η is given by $N(\widetilde{A}_\eta^{-1}\widetilde{B}_\eta, \widetilde{A}_\eta^{-1})$, where

$$\widetilde{A}_\eta = \kappa^{-2} C(h)^{-1} + \sigma^{-2} \mathrm{diag}(u_1^{-z_1}, \ldots, u_n^{-z_n}) \quad \text{and} \quad \widetilde{B}_\eta = (Y - X\beta)/\sigma^2.$$

A similar sampling strategy can be used for the two-component mixture of a normal and t-distribution with 1/2 degrees of freedom (denoted by MT), as adopted in the simulation study in Section 4. We employ the conjugate inverse gamma prior IG(1, 1) for κ^2, and a uniform prior, $U(0, h_M)$, for h, where h_M is the median of all the pairwise distances of sampling locations. The random-walk Metropolis-Hastings algorithm can be used for sampling from the full conditional distribution of h.

As an exploratory analysis, we first apply model (10) with normal error, $\varepsilon_i \sim N(0, \sigma^2)$, and computed the standardized residuals by using the posterior mean of the model parameters to visualize the potential outliers. The computed residuals are shown in the left panel of Figure 3. Although the normal error model is sensitive to outliers, there are still large residuals seen in the figure, which implies the extremity of the outliers in this dataset. In the proposed error distribution, the existence of extreme outliers is implied by the posterior of the mixture weight s, that is, the proportion of the super heavy-tailed distribution in the finite mixture. The trace plot of the posterior samples of the mixture weight s under the N-LPMN model is presented in

the right panel of Figure 3. As all the sampled values are bounded away from 0, the outliers shown in the left panel are likely to be explained by the super heavy-tailed component of the mixture. As a prior sensitivity analysis, we apply more informative priors, Beta$(1, 5)$ and Beta$(1, 9)$, in addition to the default prior $s \sim$ Beta$(1, 1)$. The posteriors computed with the three beta priors are almost identical.

The estimated spatial effects, $g(t_i)$, under the N-LPMN and normal models, are presented in Figure 4. The N-LPMN model produces spatially smoothed estimates, whereas the estimates of the normal model are volatile across the sampling area. This finding also evidences the effect of outliers on the posterior inference for the regression coefficients or, in this example, the random intercept terms.

The posterior means and 95% credible intervals of the regression coefficients based on the three methods are shown in Figure 5. This shows that the results of the normal error model are quite different from those of the MT and N-LPMN distributions. The difference in estimates becomes visually clear, especially for the significant covariates, if we define the significance in the sense that the 95% credible intervals do not contain zero, like the result of proneness/sensitivity to the representative outliers observed in Figure 3. The difference between the posteriors of the N-LPMN and MT models exists, but is not as visually clear as the difference from the normal error model.

Finally, we compute the deviance information criterion (Spiegelhalter et al., 2002, 2003; Lunn et al., 2013) of the three models. The obtained values are 2628 for the normal error model, 2339 for the MT error model, and 2325 for the proposed N-LPMN error model. This shows the best fit of the N-LPMN error model to the data using this criterion.

5.2. Diabetes data

We consider another famous dataset known as Diabetes data (Efron et al., 2004). The data contains information of 442 individuals and 10 covariates regarding the personal information and related medical information of the individuals. We consider the same formulation of linear regression model as in Efron et al. (2004); the set of predictors consists of the original 10 variables, 45 interactions, and 9 squared values, which results in $p = 64$ predictors in the model. For this dataset, the regression models with the horseshoe prior and three error distributions (N, N-LPMN, and MT) presented in Section 5.1 are adopted.

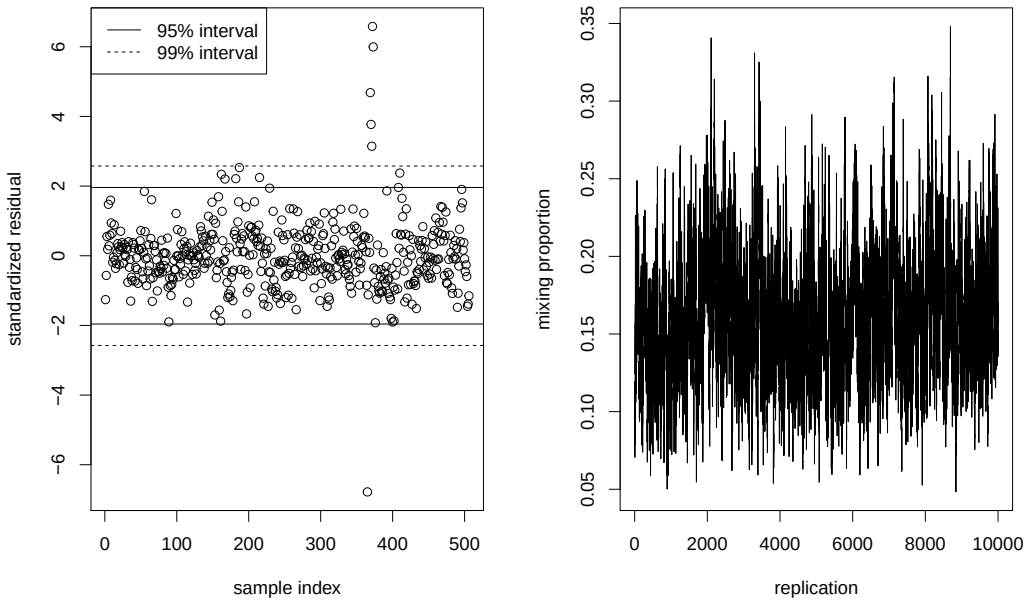

Figure 3: Standardized residuals (left) and trace plot of s (mixing proportion) in the proposed N-LPMN distribution (right), obtained form the Boston housing data. The posterior mean and the 95% credible interval of s are 0.160 and $(0.087, 0.249)$, respectively.

Similar to the analysis of the Boston housing data, we check the standardized residuals computed under the standard linear regression model. The result is presented in the left panel of Figure 6. A few outliers are confirmed in the dataset, as most of the residuals are contained in the 99% interval, which strongly supports the standard normal assumption in this example. The right panel of Figure 6 shows the trace plot of the posterior samples of mixture s under the N-LPMN distribution. All the sampled values are very close to zero, implying that most error terms should be generated from the first component of the mixture, that is, the standard normal distribution. In this case, the super heavy-tailed component might be regarded "redundant" for this dataset. The same sensitivity analysis on the choice of priors is done for s as in the previous section; however, we find no significant change in the results.

To see the possible inefficiency of using the N-LPMN models for the

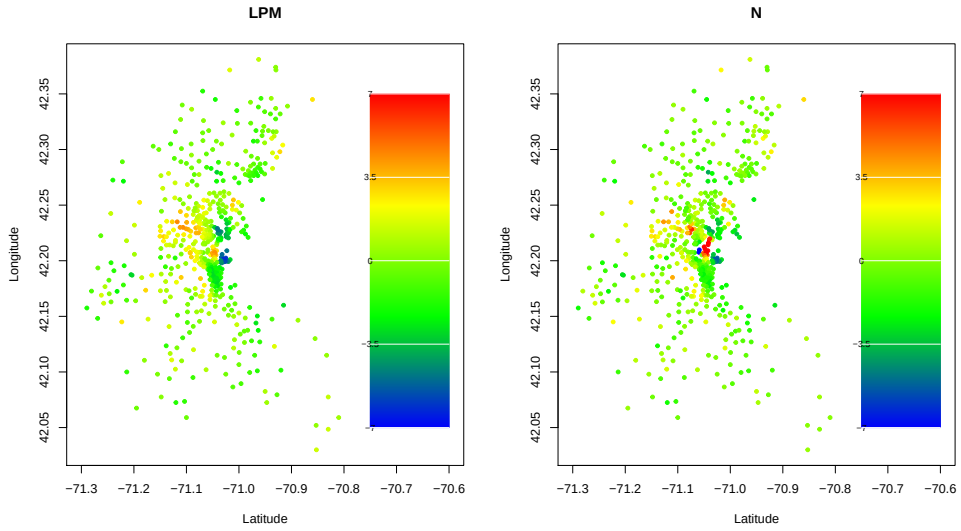

Figure 4: Posterior means of the spatial effects based on the N-LPMN and the normal (N) distribution.

dataset without outliers, the posterior means and 95% credible intervals of the regression coefficients are reported in Figure 7. The results of the three models are comparable; the predictors selected by significance are almost the same for the three models. The only notable difference is that the credible intervals produced by the t-distribution model are slightly larger than those of the other two methods. This indicates a loss of efficiency when using the t-distribution method wihtout outliers, as confirmed by the simulation results in Section 4. In contrast, the difference in the credible intervals of the Gaussian and N-LPMN models is hardly visible in the figure. That is, even if no outlier exists, the efficiency loss in the estimation under the N-LPMN model is minimal.

We also compute the deviance information criterion for the three models. The obtained values are 4794 for the normal error model and 4795 for both the MT and N-LPMN error models, which shows a comparable fit of the three models.

6. Discussions

While the focus of this research is on the inference of the regression coefficients and scale parameters, it is of great interest to employ predictive

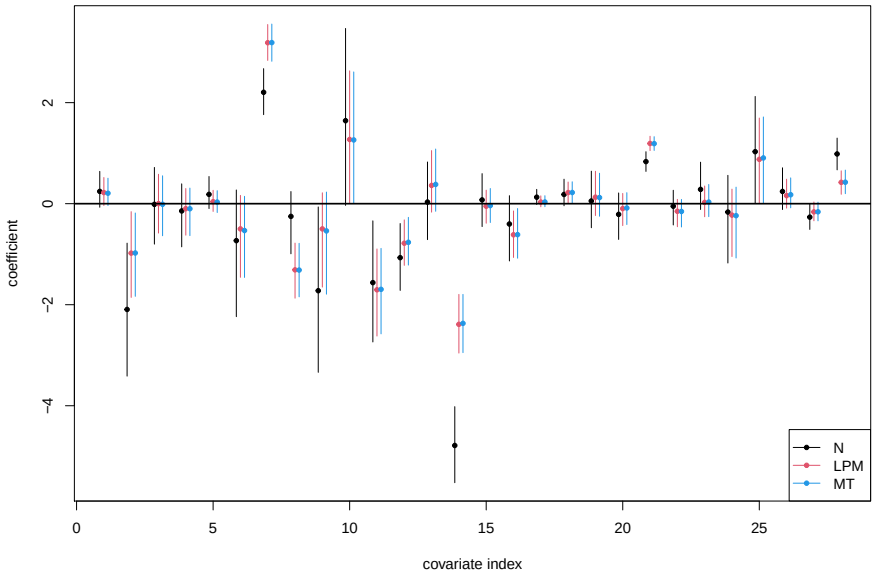

Figure 5: Posterior means and 95% credible intervals of the regression coefficients in the normal regression with normal distribution error (N), the proposed N-LPMN distribution, and the two-component mixture of normal and t-distribution with 1/2 degrees of freedom (MT), applied to the Boston housing data.

analysis based on the proposed model. Because the H-distribution, as well as many log-regularly varying distributions, is too heavily-tailed to have finite moments, posterior predictive moments under the N-LPMN models do not exist. In practice, it is common to have predictive distributions with no finite moments (West, 2020), and it is worth investigating the predictive properties under the N-LPMN models, especially regarding the impact of the super heavy tails on predictive uncertainty.

The proposed method is not limited to the analysis of the linear regression models but can be immediately customized for any models that are conditionally Gaussian, as we perform in the analysis of the random intercept model in Section 4.3 and the spatially varying intercept model in Section 5.1. Other examples include graphical models and dynamic linear models, which can be topics for promising future research. The efficient posterior computation algorithm presented in this research can be used for these highly structured models. It can also be employed when utilizing the hierar-

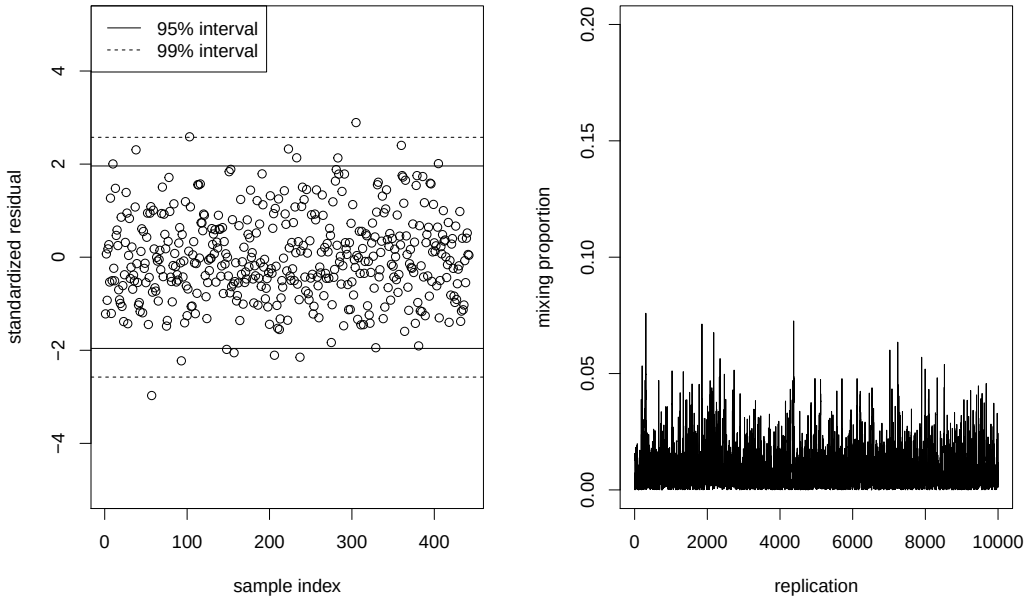

Figure 6: Standardized residuals (left) and trace plot of s (mixing proportion) in the proposed N-LPMN distribution (right), obtained form the Diabetes data. The posterior mean and the 95% credible interval of s are 0.008 and $(0.000, 0.032)$, respectively.

chical representation of the proposed error distribution. Similar theoretical robustness properties might also be confirmed for these models.

Finally, we note that the assumption (A.1) in Theorem 1 misses the high-dimensional regression with a small sample size ($n < p$), which means that posterior robustness is not necessarily achieved in this challenging situation. Therefore, substantial work is required to develop the theory and methodology for "robust high-dimensional regression," which we present as an interesting future research topic.

Acknowledgement

This work is supported by the Japan Society for the Promotion of Science (grant numbers: 20J10427, 17K17659, and 18K12757).

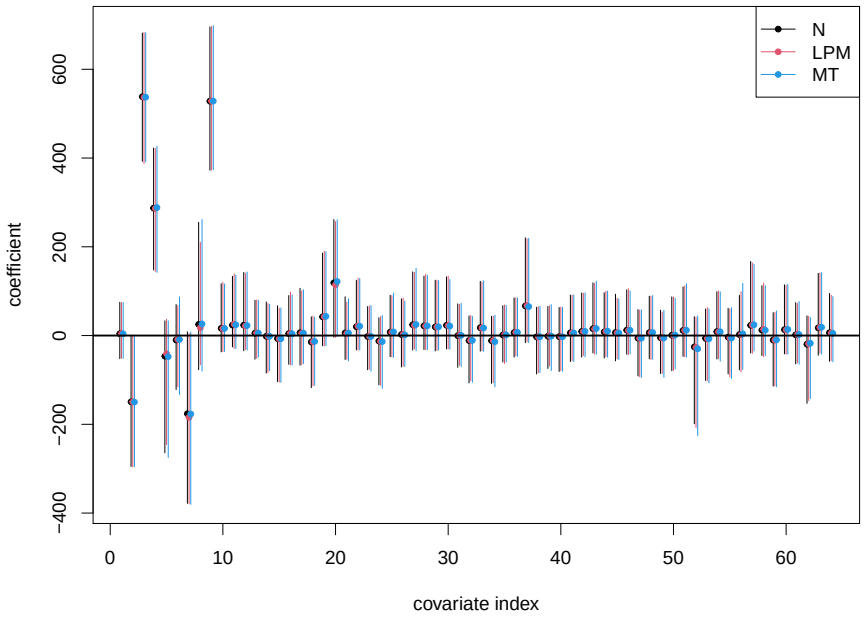

Figure 7: Posterior means and 95% credible intervals of the regression coefficients in the normal regression with normal distribution error (N), the proposed N-LPMN distribution, and the two-component mixture of t-distribution (MT) with 1/2 degrees of freedom, applied to the Diabetes data.

Supplementary Material

The proofs of all the propositions and theorems, and additional simulation results are provided in the online supplementary material.

References

Abraham, B. and Box, G. E. P. (1978). Linear models and spurious observations. *Journal of the Royal Statistical Society: Series C*, **27**, 131–138.

Andrade, J. A. A. and O'Hagan, A. (2006). Bayesian robustness modeling using regularly varying distributions. *Bayesian Analysis*, **1**, 169–188.

Andrade, J. A. A. and O'Hagan, A. (2011). Bayesian robustness modelling

of location and scale parameters. *Scandinavian Journal of Statistics*, **38**, 691–711.

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

Carter, C.K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2009). Handling Sparsity via the Horseshoe. In *AISTATS*, Volume 5, 73–80.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

Cormann, U. and Reiss, R. D. (2009). Generalizing the Pareto to the log-Pareto model and statistical inference. *Extremes*, **12**, 93–105.

Desgagné, A. (2015). Robustness to outliers in location–scale parameter model using log-regularly varying distributions. *The Annals of Statistics*, **43**, 1568–1595.

Desgagné, A. (2021). Efficient and robust estimation of regression and scale parameters, with outlier detection. *Computational Statistics & Data Analysis*, **155**, 107–114.

Desgagné, A. and Gagnon, P. (2019). Bayesian robustness to outliers in linear regression and ratio estimation. *Brazilian Journal of Probability and Statistics*, **33**, 205–221.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**, 407–499.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Gagnon, P., Desgagne, P. and Bedard, M. (2019). A New Bayesian Approach to Robustness Against Outliers in Linear Regression. *Bayesian Analysis*, **15**, 389–414.

Gelfand, A. E., Kim, H., Sirmans, C. F. and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Associatio*, **98**, 387–396.

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

Griffin, J.E. and Brown, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171–188.

Hamura, Y., Irie, K. and Sugasawa, S. (2019). On Global-local Shrinkage Priors for Count Data. *arXiv preprint arXiv:1907.01333* .

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics & Management*, **5**, 81–102.

Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. J. (2013). The BUGS book. A Practical Introduction to Bayesian Analysis. *, Chapman Hall, London*.

O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society: Series B*, **41**, 358–367.

O'Hagan, A. and Pericchi, L. (2012). Bayesian heavy-tailed models and conflict resolution: A review. *Brazilian Journal of Probability and Statistics*, **26**, 372–401.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.

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

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2020). Bayesian measures of model complexity and fit (with discussion). *Journal of the royal statistical society: Series B (statistical methodology*, **64**, 583–639.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). WinBUGS user manual, version 1.4. *Cambridge:Medical Research Council Biostatistics Unit*.

Tak, H., Ellis, J.A and Ghosh, S.K. (2019). Robust and accurate inference via a mixture of gaussian and student's t Errors. *Journal of Computational and Graphical Statistics*, **28**, 415–426.

van Dyk, D.A. and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, **103**, 790–796.

Verbeke, G. and Molenberghs, G. (2006). *Linear Mixed Models for Longitudinal Data*. Springer Science & Business Media.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, **46**, 431–439.

West, M. (1997). Modelling and robustness issues in Bayesian time series analysis (with discussion). In *Bayesian Robustness*, 231–252. Institute for Mathematical Statistics.

West, M. (2020). Bayesian decision analysis and constrained forecasting. *arXiv preprint arXiv:2007.11037*