Bayesian Analysis (2020) **15**, Number 2, pp. 389–414

A New Bayesian Approach to Robustness Against Outliers in Linear Regression

Philippe Gagnon*, Alain Desgagné†, and Mylène Bédard‡

Abstract. Linear regression is ubiquitous in statistical analysis. It is well understood that conflicting sources of information may contaminate the inference when the classical normality of errors is assumed. The contamination caused by the light normal tails follows from an undesirable effect: the posterior concentrates in an area in between the different sources with a large enough scaling to incorporate them all. The theory of conflict resolution in Bayesian statistics (O'Hagan and Pericchi (2012)) recommends to address this problem by limiting the impact of outliers to obtain conclusions consistent with the bulk of the data. In this paper, we propose a model with super heavy-tailed errors to achieve this. We prove that it is wholly robust, meaning that the impact of outliers gradually vanishes as they move further and further away from the general trend. The super heavy-tailed density is similar to the normal outside of the tails, which gives rise to an efficient estimation procedure. In addition, estimates are easily computed. This is highlighted via a detailed user guide, where all steps are explained through a simulated case study. The performance is shown using simulation. All required code is given.

Keywords: ANOVA, ANCOVA, built-in robustness, maximum likelihood estimation, super heavy-tailed distributions, variable selection, whole robustness.

MSC 2010 subject classifications: Primary 62F35; secondary 62J05.

1 Introduction

The distribution most commonly assumed on the error term in the linear regression model $Y = \mathbf{x}^T\boldsymbol{\beta}+\epsilon$ is without a doubt a normal, denoted $\epsilon/\sigma \sim \mathcal{N}(0,1)$. Estimating the regression coefficient vector $\boldsymbol{\beta}$ is in this case equivalent to using ordinary least squares (OLS) method, whether Bayesian (setting the usual noninformative prior on $\boldsymbol{\beta}$) or maximum likelihood estimates (MLE) are computed. Given the remarkable properties of OLS (under certain conditions) such as minimum variance among unbiased estimators, the normal model is often considered as a benchmark in terms of efficiency in the absence of outliers. However, it is well-known that resulting inferences are very sensitive to conflicting sources of information. From a Bayesian perspective, these conflicting sources may represent the prior or outliers; we focus on the latter in this paper.

*Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, United Kingdom, philippe.gagnon@stats.ox.ac.uk

†Département de mathématiques, Université du Québec à Montréal, C.P. 8888, Succursale Centre-ville, Montréal, QC, H3C 3P8, Canada, desgagne.alain@uqam.ca

‡Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montréal, QC, H3C 3J7, Canada, bedard@dms.umontreal.ca

 https://doi.org/10.1214/19-BA1157

Box and Tiao (1968) were the first to propose a Bayesian solution. They suggested to let the error term be distributed as a mixture of two normals: one component for the nonoutliers and the other one, with a larger variance, for the outliers. This approach has been generalised by West (1984) who modelled errors with heavy-tailed distributions constructed as scale mixtures of normals, which include the Student distribution. A different robust Bayesian approach was introduced by Peña et al. (2009). From a frequentist perspective, several methods have also been proposed, e.g., the M- (Huber (1973)), MM- (Yohai (1987)), S- (Rousseeuw and Yohai (1984)), least trimmed squares (LTS, Rousseeuw (1985)), and robust and efficient weighted least-square (REWLSE, Gervini and Yohai (2002)) estimators.

The most popular Bayesian solution is modelling using the Student, a consequence of the simplicity of the strategy, the rationale behind it (giving higher probabilities to extreme values), and the required computations. The latter follows from the scale mixture representation of the Student that leads to a normal conditional distribution for Y given $\boldsymbol{\beta}$, σ and a latent variable, which in turn allows a straightforward implementation of the Gibbs sampler (Geman and Geman (1984)). This method took over that of Box and Tiao (1968) because the latter is such that the conditional distribution is a mixture of normals and requires to "complete" the data with auxiliary variables to implement the Gibbs sampler. This may make computations much more arduous. On the frequentist side, the most popular method to gain in robustness is arguably the MM-estimator.

Protection against outliers always comes at a price: a loss of efficiency when the observations are normally distributed. The best robust alternatives manages to offer a large protection at a low premium. This is especially true for the estimation of $\boldsymbol{\beta}$. In this regard, a new method can hardly do better; in fact matching their performance is quite an achievement. However, the performance of the existing robust approaches with respect to σ is far less optimal.

The main objective of this paper is to propose a solution that yields gold standard performance, namely a large protection at a low premium, for the estimation of both $\boldsymbol{\beta}$ and σ. The importance of good estimation for σ, in the absence or presence of outliers, should not be overlooked. This parameter plays a crucial role every time an assessment has to be made about uncertainty around the regression coefficients (credible intervals, hypothesis testing, and so on). The performance of the proposed approach, combined with its simplicity, will allow to offer an appealing Bayesian alternative to the Student model.

The first step towards the objective is indeed to employ a strategy as simple as that of West (1984), that is, to assume a distribution on the error term that accommodates for the eventual presence of outliers without being a mixture. Our approach differs in that the density has a slower tail decay. It is based on the work of Desgagné (2015) about robust modelling of location and scale parameters. The author proposed to use a super heavy-tailed distribution belonging to the family of log-regularly varying distributions (LRVD) — with tails behaving like $|z|^{-1}(\log|z|)^{-\theta}$ — to achieve whole robustness for both parameters. The idea of using heavier tails than the Student came after the work of Andrade and O'Hagan (2011) who, in the location-scale framework, achieved only partial robustness for the scale by modelling with polynomial tails. As mentioned by

West (1984), an outlying observation is accommodated if the posterior distribution converges to that excluding the outlier as this one tends to infinity, which corresponds to our definition of whole robustness. In contrast, partial robustness translates into a significant (but limited as the outliers approach plus or minus infinity) impact on the estimation of the parameter.

The second step towards the objective is therefore to generalise the results of Desgagné (2015) to linear regression. In fact, it is a generalisation of the results of Desgagné and Gagnon (2019), which are essentially an application of those of Desgagné (2015) in simple linear regression through the origin for robust estimation of ratios. This second step represents our key theoretical contribution. We provide two sufficient conditions that lead to whole robustness. The first one is to assume a super-heavy tailed distribution on the error. The other specifies the breakdown point, which tends to the optimal value of 0.5 as the sample size goes to infinity. The validity of our robust method is thus supported by theoretical results. While these are similar to those of Desgagné and Gagnon (2019), a more sophisticated proof technique is required given that the location parameter of the conditional distribution of Y is now an inner product of a known vector and $\boldsymbol{\beta}$ containing p unknown parameters. Throughout the paper, we focus on continuous explanatory variables to simplify explanation and notation. The results are nonetheless valid in ANOVA and ANCOVA (analyses of variance and covariance), and for variable selection where joint posteriors of models and parameters are considered. The corresponding sufficient conditions are given as remarks after the theoretical results. The price to pay to achieve whole robustness for all parameters is that the use of super heavy-tailed distributions prevents us from obtaining normal conditional distributions. There is therefore a computational cost, in the sense that we cannot implement a Gibbs sampler; it will however be noticed that easy-to-use samplers can be used, which makes the cost negligible.

The third and final step towards the objective is to carefully select the super heavy-tailed distribution in the wholly robust model. To achieve this, we start with the premise that applied statisticians are satisfied with the normal model in the absence of outliers and we specifically design a robust solution from that. We set the distribution of the error as a log-Pareto-tailed normal (LPTN), a super heavy-tailed distribution introduced by Desgagné (2015). Its density exactly matches the standard normal on the central part having a mass of ρ. The parameter ρ is thus the single one to be chosen by the user, and is typically set to a value between 0.80 and 0.98. The resulting model produces robust estimates exhibiting a similar behaviour to OLS in the absence of outliers, where the trade-off between high degree of similarity with OLS and high degree of robustness is controlled through ρ. The model has built-in robustness that resolves conflict in a sensitive way (see Figure 1). It completely considers the nonoutliers (from 30 to 32.5 in Figure 1), essentially excludes the observations that are clearly outlying (beyond 38 in Figure 1), and between these two clear cases, contains and bounds their impact. The first two cases correspond to the strategy commonly applied in practice, where an observation is either kept or discarded. In the last case, the method reflects that in the gray area there is a level of uncertainty about the fact that those observations really are outliers or not. Our main practical contribution is therefore to provide an efficient and robust model that automatically deals with this type of uncertainty, which is especially valuable in high-dimensional problems and when several analyses have to be performed.

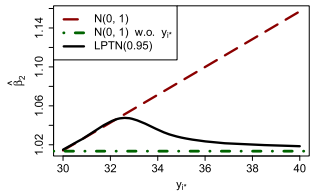

Figure 1: Posterior mean of the slope in a simple linear regression as an observation $y_{i^*} \to \infty$.

This rest of the article is organised as follows. The linear regression model is detailed in Section 2.1, the LRVD family is presented in Section 2.2 and the theoretical results are provided in Section 2.3. More practically, efficient and robust regression is investigated in Section 3. The LPTN distribution is first presented in Section 3.1. A discussion about efficiency of the robust model with LPTN errors is provided in Section 3.2. Practical details of our approach are addressed in Section 3.3 through a simulated case study on the modelling of house market values. Numerical methods such as Markov chain Monte Carlo (MCMC) are discussed for the computation of different posterior quantities: means, medians, credible intervals, prediction of future observations and hypothesis testing via Bayes factors. A powerful tool for outlier identification is also proposed. In Section 3.4, a simulation study is conducted to compare the performance of our approach with different Bayesian alternatives. Note that even though our approach is Bayesian, it is possible to use it in a frequentist setting through maximum a posteriori probability (MAP) estimates, which correspond to MLE when the prior is set to 1. We thus also include in our study the frequentist methods mentioned above.

2 Conflict Resolution in Linear Regression via LRVD

We henceforth assume that f is a strictly positive continuous probability density function (PDF) on $\mathbb{R}$ that is symmetric with respect to the origin, for which all parameters are known and such that there exists a threshold above which $g(z) = zf(z)$ is monotonic. Examples of such PDF are the normal, logistic, Laplace, Student (with prespecified degrees of freedom) and the LPTN (see Section 3.1).

2.1 Linear Regression Model

(i) Let $Y_1, \ldots, Y_n \in \mathbb{R}$ be n random variables representing data points from the dependent variable and $\mathbf{x}_1^T := (1, x_{12}, \ldots, x_{1p}), \ldots, \mathbf{x}_n^T := (1, x_{n2}, \ldots, x_{np})$ be n vectors of observations from the explanatory variables, where $p \in \{2, 3, \ldots\}$, $n \geq p+1$ and $x_{ij} \in \mathbb{R}$ are assumed to be known. As mentioned in the introduction, we focus on the situation where all explanatory variables are continuous. The linear regression model is given by

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where the n random variables $\epsilon_1, \ldots, \epsilon_n \in \mathbb{R}$ and the p-dimensional random variable $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ represent the errors and the vector containing the regression coefficients, respectively. These $n+1$ random variables are conditionally independent given $\sigma > 0$, a scale parameter, with a conditional density for ϵ_i given by

$$\epsilon_i \mid \boldsymbol{\beta}, \sigma \overset{\mathcal{D}}{=} \epsilon_i \mid \sigma \overset{\mathcal{D}}{\sim} (1/\sigma) f\left(\epsilon_i/\sigma\right), \quad i = 1, \ldots, n.$$

(ii) We assume that the joint prior density of $\boldsymbol{\beta}$ and σ, denoted $\pi(\boldsymbol{\beta}, \sigma)$, is bounded by $\max(C, \sigma^{-1}C)$, where $C > 0$ can be any constant.

A large variety of priors fits within the structure assumed in (ii). This is the case for noninformative priors such as $\pi(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$ and $\pi(\boldsymbol{\beta}, \sigma) \propto 1$, and practically all proper densities. Informative priors shall however be used with caution, especially when they translate into light tailed densities. They may indeed contaminate the inference if they are in conflict with the information carried by the data. Establishing the conditions that guarantee robustness to informative priors in linear regression is not trivial.

We study robustness of the estimation of $\boldsymbol{\beta}$ and σ in the presence of outliers. In this paper, an observation $(\mathbf{x}_i, y_i)$ is considered as an outlier if its error $\epsilon_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$ is relatively far from 0, where $\boldsymbol{\beta}$ defines the probable hyperplanes for the bulk of the data. Note that robustness against outlying errors is a different concept than robustness against outlying $\mathbf{x}_i$ or y_i. They are generally equivalent though, except for the unusual case where an observation is outlying in $\mathbf{x}_i$ and y_i but still manages to lie in the general trend, and consequently, be a nonoutlier in error. From a theoretical perspective, we study the asymptotic behaviour in the sense that we let outliers' errors ϵ_i approach $+\infty$ or $-\infty$. Our strategy to mathematically represent this situation is to let their y_i approach $+\infty$ or $-\infty$ while their vector $\mathbf{x}_i$ remains fixed. We thus specify a particular path along which the outliers move away from the general trend.

We assume that each outlier goes to $-\infty$ or $+\infty$ at its own specific rate, to the extent that the ratio of two outliers is bounded. More precisely, we assume that

$$y_i = a_i + b_i\omega, \tag{2.2}$$

for $i = 1, \ldots, n$, where $a_i, b_i \in \mathbb{R}$ are constants such that $b_i = 0$ if the point is a nonoutlier and $b_i \neq 0$ if it is an outlier, and then, we let $\omega \to \infty$. We mathematically distinguish the outliers from the nonoutliers through the following. Among the n observations $(y_1, \ldots, y_n) =: \mathbf{y_n}$, we assume that k of them form a group of nonoutlying observations, that we denote $\mathbf{y_k}$, while $\ell = n - k$ of them are considered as outliers. For $i = 1, \ldots, n$, we define the binary functions k_i and ℓ_i as follows: if y_i is a nonoutlying value $k_i = 1$, and if it is an outlier $\ell_i = 1$. These functions take the value of 0 otherwise. Therefore, we have $k_i + \ell_i = 1$ for $i = 1, \ldots, n$, with $\sum_{i=1}^n k_i = k$, and $\sum_{i=1}^n \ell_i = \ell$.

Let the joint posterior density of $\boldsymbol{\beta}$ and σ be denoted by $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n})$ and the marginal density of $(Y_1, \ldots, Y_n)$ be denoted by $m(\mathbf{y_n})$, where

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n}) = [m(\mathbf{y_n})]^{-1}\pi(\boldsymbol{\beta}, \sigma)\prod_{i=1}^{n}(1/\sigma)f((y_i - \mathbf{x}_i^T\boldsymbol{\beta})/\sigma), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0. \tag{2.3}$$

Let the joint posterior density of $\boldsymbol{\beta}$ and σ arising from the nonoutlying observations only be denoted by $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k})$ and the corresponding marginal density be denoted by $m(\mathbf{y_k})$, where

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k}) = [m(\mathbf{y_k})]^{-1} \pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^{n} \left[(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)\right]^{k_i}, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0.$$

Proposition 2.1 (Tail behaviour of the posteriors)**.**

(i) *If $n > p + 1$, the density $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n})$ is proper.*

(ii) *If $k > p + 1$ (stronger than $n > p + 1$), the density $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k})$ is also proper.*

(iii) *If $n > p + 1 + M$, then $\mathbb{E}[\beta_j^M \mid \mathbf{y_n}]$ for any $j \in \{1, \ldots, p\}$ and $\mathbb{E}[\sigma^M \mid \mathbf{y_n}]$ exist.*

(iv) *If $k > p + 1 + M$, then $\mathbb{E}[\beta_j^M \mid \mathbf{y_k}]$ for any $j \in \{1, \ldots, p\}$ and $\mathbb{E}[\sigma^M \mid \mathbf{y_k}]$ exist.*

Proof. See Section 5. □

Remark 2.1. *When any type of explanatory variables is considered (continuous, discrete as in ANOVA or a mix of both as in ANCOVA), the densities are proper if we additionally assume that the design matrix (comprised of n or k observations) has full rank. In variable selection, when the joint posterior of the models and parameters is considered, this joint posterior is proper if the assumptions are verified for the "complete" model (the model with all variables). The assumptions are more technical for the moments and are not provided here. We essentially need enough of "different" $\mathbf{x}_i$ vectors. In the proof, it is made clear what is required.*

2.2 Log-Regularly Varying Distributions

We now provide an overview of the class of log-regularly varying functions (LRVF), as introduced in Desgagné (2013) and Desgagné (2015), following the idea of regularly varying functions developed by Karamata (1930). They form an interesting class of functions with useful properties for robustness.

Definition 2.1 (LRVF)**.** *We say that a measurable function g is log-regularly varying at ∞ with index $\theta \in \mathbb{R}$, written $g \in L_\theta(\infty)$, if*

$$\lim_{z \to \infty} g(z^\nu)/g(z) = \nu^{-\theta},$$

uniformly in any set $\nu \in [1/\eta, \eta]$ (for any $\eta \geq 1$). If $\theta = 0$, g is said to be log-slowly varying at ∞.

In Desgagné (2015), it is shown that Definition 2.1 is equivalent to the following: there exists a constant $A > 1$ and a function $s \in L_0(\infty)$ such that for $z \geq A$, g can be written as

$$g(z) = (\log z)^{-\theta} s(z).$$

Examples of LRVF are $g(z) = (\log z)^{-\theta}$ (with $s(z) = 1$) and $g(z) = (\log z)^{-\theta} \log(\log z)$.

Definition 2.2 (LRVD)**.** *A random variable Z and its distribution are said to be log-regularly varying with index $\theta \geq 1$ if their density f is such that $zf(z) \in L_\theta(\infty)$.*

Definition 2.2 implies that any density f with tails behaving like $|z|^{-1}(\log|z|)^{-\theta}$ with $\theta > 1$ is a LRVD. Some examples like the LPTN distribution are given in Desgagné (2015). The most important property of this class of distributions follows from Definition 2.1: the asymptotic location-scale invariance of their density, as stated in Proposition 2.2.

Proposition 2.2 (Location-scale invariance)**.** *If $zf(z) \in L_\theta(\infty)$, then we have*

$$(1/\sigma)f((z-\mu)/\sigma)/f(z) \to 1 \text{ as } z \to \infty,$$

uniformly on $(\mu, \sigma) \in [-\vartheta, \vartheta] \times [1/\eta, \eta]$, for any $\vartheta \geq 0$ and $\eta \geq 1$.

Proof. See Desgagné (2015). □

Proposition 2.2 essentially implies that the conditional density of an outlier $(1/\sigma)$ $f((y - \mathbf{x}^T\boldsymbol{\beta})/\sigma)$ asymptotically behaves like $f(y)$ as $y \to \infty$. The densities of the outliers at the numerator of posterior densities cancel each other out with those at the denominator in the marginal, provided that the integral can be interchanged with the limit. This is the idea of the proof of our robustness result presented in the next section. The greatest challenge is however to prove that we can indeed interchange the limit and the integral. This part leads to the condition about the maximum number of outliers to guarantee robustness.

2.3 Resolution of Conflicts

We now present Theorem 2.1, the main theoretical contribution of this paper.

Theorem 2.1. *If*

(i) *$zf(z) \in L_\theta(\infty)$ with $\theta \geq 1$, i.e. f is a LRVD,*

(ii) $$\begin{aligned} &\ell \leq n/2 - (p - 1/2), \text{ i.e. } \#\text{outliers} \leq \text{half the sample} - (p-1/2), \\ \Leftrightarrow\ &k \geq n/2 + (p - 1/2), \text{ i.e. } \#\text{nonoutliers} \geq \text{half the sample} + (p-1/2), \\ \Leftrightarrow\ &k - \ell \geq 2(p-1/2), \quad \text{i.e. } \#\text{nonoutliers} - \#\text{outliers} \geq 2(p-1/2), \end{aligned}$$

then, as $\omega \to \infty$ (where ω is defined in (2.2)), we obtain the following results:

(a)

$$\frac{m(\mathbf{y_n})}{\prod_{i=1}^{n}[f(y_i)]^{\ell_i}} \to m(\mathbf{y_k}),$$

(b)

$$\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n}) \to \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k}),$$

uniformly on $(\boldsymbol{\beta}, \sigma) \in [-\vartheta, \vartheta]^p \times [1/\eta, \eta]$, for any $\vartheta \geq 0$ and $\eta \geq 1$,

(c)

$$\boldsymbol{\beta}, \sigma \mid \mathbf{y_n} \xrightarrow{\mathcal{D}} \boldsymbol{\beta}, \sigma \mid \mathbf{y_k},$$

and in particular

$$\beta_j \mid \mathbf{y_n} \xrightarrow{\mathcal{D}} \beta_j \mid \mathbf{y_k},\ j = 1, \ldots, p, \qquad \textit{and} \qquad \sigma \mid \mathbf{y_n} \xrightarrow{\mathcal{D}} \sigma \mid \mathbf{y_k},$$

(d) *if additionally* $k \geq n/2 + (p - 1/2) + M$, *then*

$$\mathbb{E}[\beta_j^M \mid \mathbf{y_n}] \to \mathbb{E}[\beta_j^M \mid \mathbf{y_k}],\ j = 1, \ldots, p, \qquad \textit{and} \qquad \mathbb{E}[\sigma^M \mid \mathbf{y_n}] \to \mathbb{E}[\sigma^M \mid \mathbf{y_k}].$$

Proof. See Section 5. □

The two sufficient conditions of Theorem 2.1 are remarkably simple. Condition (i) indicates that modelling must be performed using a super heavy-tailed density f, more precisely using a LRVD, e.g. a LPTN as proposed. Condition (ii) gives in fact the breakdown point, generally defined as the proportion of outliers (ℓ/n) that an estimator can handle. We have $\ell/n \leq 1/2 - (p - 1/2)/n$, which translates into a breakdown point of 50% as $n \to \infty$ (for fixed p), usually considered as the maximum and best desired value. Condition (ii) is thus generally satisfied in practice.

Results (a) to (d) are different representations of whole robustness. Essentially, the posterior inference arising from the whole sample converges towards the posterior inference based on the nonoutliers only. The impact of outliers then gradually vanishes as they approach plus or minus infinity.

In Result (a), the asymptotic behaviour of the marginal $m(\mathbf{y_n})$ is described. This result is used in Section 3.3 to assess robustness of Bayes factors for testing $H_0 : \beta_i = 0$ versus $H_0 : \beta_i \neq 0$ (when $i \geq 2$). Result (a) is in fact the centrepiece of Theorem 1; its demonstration requires considerable work, and leads relatively easily to the other results of the theorem.

The convergence of the posterior density in Result (b) enables to assess that the MAP estimates of $\boldsymbol{\beta}$ and σ are wholly robust. Given that these estimators correspond to the MLE when the prior is proportional to 1, the frequentist estimates are, as a result, also wholly robust. This allows establishing a connection between Bayesian and frequentist robustness.

Result (c) indicates that any estimation of $\boldsymbol{\beta}$ and σ based on posterior quantiles (e.g. using posterior medians and Bayesian credible intervals) is robust to outliers. Note that in fact we obtain the stronger result of L^1 convergence:

$$\int_0^\infty \int_{\mathbb{R}^p} \left| \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n}) - \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k}) \right| d\boldsymbol{\beta}\, d\sigma \to 0,$$

which in turn implies that $\mathbb{P}(\boldsymbol{\beta}, \sigma \in E \mid \mathbf{y_n}) \to \mathbb{P}(\boldsymbol{\beta}, \sigma \in E \mid \mathbf{y_k})$ as $\omega \to \infty$, uniformly for all sets $E \subset \mathbb{R}^p \times \mathbb{R}^+$, a slightly stronger than convergence in distribution given in Result (c) which requires only pointwise convergence.

Posterior expectations are wholly robust as well, as indicated by Result (d). It is interesting to notice that all these results guarantee the robustness of a variety of Bayes estimators.

Remark 2.2. *When any type of explanatory variables is considered, the same results as in Theorem 2.1 hold under the following additional assumption: it is possible to choose* $n/2 + (p - 1/2)$ *(or* $n/2 + (p - 1/2) + M$*) nonoutliers — the required number of nonoutliers depending on which results we target (Results (a) to (c) or Results (a) to (d)) — that have p-wise linearly independent* $\mathbf{x}_i$ *vectors. This means that any p vectors* $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}$ *among the chosen subgroup must be linearly independent. In variable selection, the convergence of the joint posterior of the models and their parameters, and of the expectations, hold if the assumptions are verified for the complete model.*

Remark 2.3. *We prove that modelling with f having tails behaving like* $|z|^{-1}(\log|z|)^{-\theta}$ *is sufficient to obtain the results in Theorem 2.1. It seems "almost" necessary because, on one hand, a tail behaviour of* z^{-2} *(corresponding to a Student density) is not sufficient, and on the other hand,* $|z|^{-1}$ *is not integrable.*

3 Efficient and Robust Regression Using LPTN

In Section 2.3, we stated theoretical results which essentially indicate that using a LRVD for the errors ensures a high breakpoint of $1/2 - (p - 1/2)/n$ with a whole rejection of the outliers as their error goes to $+\infty$ or $-\infty$. The conflict is thus resolved and the linear regression is in agreement with the bulk of the data.

In this section, we build on these results to propose a solution in the realistic situation where a statistician satisfied with the normal model in the absence of outliers seeks protection in the eventuality of contamination by outliers. Mathematically, we consider the context where the errors come from a mixture distribution, with a normal component for the bulk of the data and another component F_0 for the outliers, that is

$$\epsilon_i/\sigma \sim \alpha\mathcal{N}(0,1) + (1-\alpha)F_0, \quad i = 1, \ldots, n, \tag{3.1}$$

where $0 < \alpha \leq 1$ represents the proportion of normal observations in the sample. We thus look for efficient estimators that perform well in the absence of outliers, that is when $\alpha = 1$ and the model is the pure normal. As mentioned in the introduction, OLS (or equivalently the normal model) is considered as the benchmark in this situation. Our efficient estimators must also be robust and perform in the presence of outliers, and this, for as many scenarios of $\alpha < 1$ and F_0 as possible.

3.1 LPTN Distribution

The solution we propose consists in assuming that the errors have a LPTN distribution with a prespecified parameter $\rho \in (2\Phi(1) - 1, 1) \approx (0.6827, 1)$, denoted LPTN($\rho$). More precisely, we still have $\epsilon_i \mid \sigma \overset{\mathcal{D}}{\sim} (1/\sigma) f(\epsilon_i/\sigma)$, but the density f is now assumed to be

$$f(z) = \begin{cases} \varphi(z) & \text{if} \quad |z| \leq \tau, \\ \varphi(\tau)\,\frac{\tau}{|z|}\left(\frac{\log\tau}{\log|z|}\right)^{\lambda+1} & \text{if} \quad |z| > \tau, \end{cases} \tag{3.2}$$

where $z \in \mathbb{R}$, and $\tau > 1$ and $\lambda > 0$ are functions of ρ with

$$\tau = \Phi^{-1}((1+\rho)/2) := \{\tau : \mathbb{P}(-\tau \leq Z \leq \tau) = \rho \text{ for } Z \overset{\mathcal{D}}{\sim} \mathcal{N}(0,1)\}, \qquad (3.3)$$
$$\lambda = 2(1-\rho)^{-1}\varphi(\tau)\,\tau\log(\tau),$$

$\varphi(\cdot)$, $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ being the PDF, cumulative distribution function (CDF) and inverse CDF of a standard normal, respectively.

The LPTN distribution was introduced by Desgagné (2015), who in fact presents a more general version than that shown here. The parameter λ that controls the tail decay was originally free and a multiplicative normalising constant $K(\rho, \lambda)$ was needed. For example, the center of the density (the area $|z| \leq \tau$) was given by $K(\rho, \lambda)\varphi(z)$. In order to pursue our efficiency objective, we set the constant to 1, which in return forces λ to be automatically set as a function of ρ. The parameter ρ, chosen by the user, thus represents the mass of the central part that exactly matches the $\mathcal{N}(0,1)$ density.

As ρ increases, f approaches the normal. An increase in ρ also implies an increase in λ and τ, which translates into a density f with lighter tails. Efficiency is also expected to increase, but robustness to decrease. A compromise has therefore to be made and it is controlled by the statistician through the parameter ρ. In other words, this parameter represents the tolerance to (bounded) impact from outliers at the benefit of efficiency when the data set is not contaminated. The user can also select its value based on prior opinion about the probable proportion of outliers, by setting it to 1 minus this proportion.

The rationale behind proposing the LPTN is thus that, in addition to exactly matching the normal density on the part with highest probability, this distribution has log-Pareto tails ensuring that our theoretical robustness result hold, and this for any value of ρ. This type of tails consequently accommodates for a large spectrum of α and F_0 in the mixture (3.1) when $\alpha < 1$ and generates efficient inference when $\alpha = 1$ as well (this latter characteristic is discussed in Section 3.2). A comparison between different LPTN densities is shown in Figure 2. Note that, as required for our theoretical results of Section 2, the LPTN distribution has a strictly positive continuous PDF on $\mathbb{R}$ that is symmetric with respect to the origin and such that $zf(z)$ is monotonic for $z > \tau$.

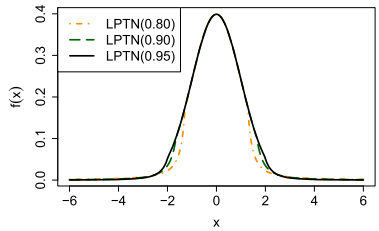

Figure 2: Densities of the LPTN(0.80), LPTN(0.90) and LPTN(0.95).

3.2 Efficiency of the LPTN Model

To theoretically study the efficiency of the LPTN Model, we consider the situation where the data are generated from a normal and evaluate the performance of the robust estimators in the asymptotic situation $n \to \infty$. We start by providing evidences that the estimators for $\boldsymbol{\beta}$ are consistent, while it depends on ρ for σ. We consider that the generative normal model has $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and $\sigma_0 > 0$ as true parameter values, and denote the associated density of one data point $g := \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta}_0, \sigma_0^2)$. Denote that associated with the LPTN model $p_{(\boldsymbol{\beta},\sigma)}(y_i) := (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)$, where f is a LPTN(ρ). In Bunke et al. (1998), it is proved that if the divergence

$$\mathrm{KL}(\boldsymbol{\beta}, \sigma) := \int \log(g(y_i)/p_{(\boldsymbol{\beta},\sigma)}(y_i))\, g(y_i)\, dy_i \tag{3.4}$$

is minimised at a unique $(\boldsymbol{\beta}^*, \sigma^*)$ and some regularity conditions are satisfied, then

$$\lim_{n\to\infty} \mathbb{E}[(\boldsymbol{\beta}, \sigma) \mid \mathbf{y_n}] = (\boldsymbol{\beta}^*, \sigma^*) \quad \text{with probability 1,}$$

where the expectation is with respect to the posterior arising from the LPTN model. This is proved through the strong consistency of the MAP.

In the supplementary material (Gagnon et al., 2019), we prove that the first derivative of (3.4) with respect to $\boldsymbol{\beta}$ equals 0 at $\boldsymbol{\beta}_0$, and this for any value of σ. While setting $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in (3.4), we show that it is minimised at σ^* which depends on ρ (see Figure 3). We also show that most of the regularity conditions in Bunke et al. (1998) are satisfied. This analysis suggests that the true values for the regression coefficients are recovered even though the LPTN model is misspecified. For σ, the closer ρ is to 1, the more similar are σ^* and σ_0. For instance, when $\rho = 0.9$, $\sigma^*/\sigma_0 = 1.03$, and beyond $\rho = 0.95$, this ratio is essentially 1.

When the data are generated from the normal model, estimators arising from it are certainly more efficient. We however numerically verified that the learning rate for the robust estimators is the same as the normal ones, suggesting that the efficiency is bounded away from 0 for all n. Some additional details are needed to rigorously prove the consistency of the Bayes estimates and to accurately conclude about efficiency.

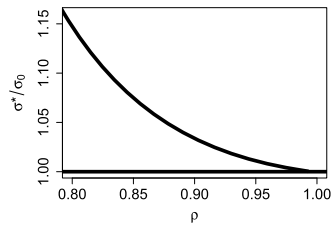

Figure 3: Minimiser of the divergence σ^* when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, as a function of ρ.

3.3 Simulated Case Study

We carry out in this section a linear regression analysis on a given data set using our robust approach and also the classical method with the normal assumption for comparison. In doing so, we address all practical considerations, resulting in a straightforward implementation by users. In this regard, all R code used to produce numerical results is provided at https://arxiv.org/abs/1612.06198, which also allows reproducing these results.

For a given city, we want to model the market value of a house in thousands of dollars using the average home value in its residential sector in thousands of dollars, the living area in square metre (sq.m.) and the land area in sq.m. We consider a simulated sample of size $n = 50$ that contains 3 outliers (it is given in detail in the provided R code). To give an overview of it, we present in Table 1 the data for Home 2 and for the outliers: Homes 1, 3 and 49.

Characteristics	**Home 2**	**Home 1**	**Home 3**	**Home 49**
Home value (in $1,000)	326	137	20	1,000
Value of the sector (in $1,000)	343	670	350	560
Living area (in sq.m)	205	149	222	269
Land area (in sq.m)	345	372	434	655

Table 1: Data from the studied sample.

Home 2 has a value of $326,000 (the sample mean is $504,900), is located in a residential sector where houses are valued at $343,000 in average (the sample mean is $508,880), has a living surface of 205 sq.m. (the sample mean is 200 sq.m) and a land of 345 sq.m. (the sample mean is 500 sq.m). Homes 1 and 3 both have aberrantly low values, while it is the opposite for Home 49. They are meant to represent a damaged house, a data entry error and an eco-friendly house, respectively.

To improve the interpretation of the linear regression, the explanatory variables are centred around their respective sample mean. Therefore, for each house, we define x_{i2} as the average value in its residential sector (in $1,000) minus 508.88, x_{i3} as the living area minus 200 and x_{i4} as the land area minus 500. Note that centring affects only the constant of the model, β_1, which can now be interpreted as the predicted value of the typical house with average features $x_{i2} = x_{i3} = x_{i4} = 0$. The model used to generate the data (except the outliers) is $Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$ with $\boldsymbol{\beta} := (508.88, 1, 1, 0.5)^T$ and $\epsilon_i \mid \sigma \overset{\mathcal{D}}{\sim} (1/\sigma) f(\epsilon_i/\sigma)$, where $f = \mathcal{N}(0,1)$ and $\sigma = 40$.

In Figure 4, we plot the dependent variable against each explanatory variable to depict their respective linear relation. The pairwise correlations between the explanatory variables are all below 0.10, suggesting that these graphs provide a fair representation of the multivariate relation. The parameters of the generative model have been set to create the expected situation in which an increase in any feature is associated with an increase in home value.

For the analysis, the density f is assumed to be a LPTN($\rho = 0.95$) for the robust model and a $\mathcal{N}(0,1)$ under the classical model. We also set $\pi(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$, the usual noninformative prior. The estimation of the parameters is done through the posterior

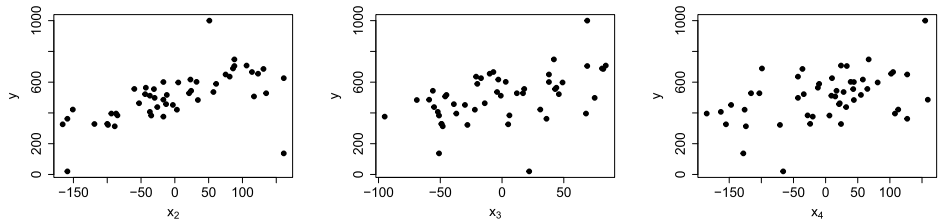

Figure 4: The dependent variable versus each of the covariates.

density as expressed in (2.3). The posterior means, medians and credible intervals are computed through a random walk Metropolis (RWM) algorithm, one of the easiest to implement Metropolis–Hastings (Metropolis et al. (1953) and Hastings (1970)) algorithms. More sophisticated methods like the Hamiltonian Monte Carlo (HMC, see, e.g., Neal (2011)) could be used given that the likelihood function is differentiable almost everywhere. The MAP and MLE are computed through optimisation procedures; we use the general-purpose `optim` function in R based on Nelder–Mead algorithm. It is of common knowledge that maximisers (MAP and MLE) may not provide a posterior summary as good as posterior means, for instance. The advantage is that they can be computed quickly. We find them particularly useful for directly giving starting points for the RWM algorithm and for conducting simulation studies as in Section 3.4.

These estimates are presented in Table 2, in which the numbers in square brackets are those based on the 47 nonoutliers only (the sample without Homes 1, 3 and 49). The lower and upper bounds of the credible intervals (CI – LB and CI – UB) are computed from the regions with highest posterior density using the `coda` package. Some interesting observations are now made. First, in the absence of outliers (results in brackets), the results of the robust LPTN model are very similar to those of the nonrobust normal model. As mentioned in Section 3.1, the LPTN(0.95) is very similar to the $\mathcal{N}(0,1)$, in fact identical except for the 5% tails. The normal model is the benchmark in terms of efficiency. All presented point estimators of $\boldsymbol{\beta}$ under the normal model indeed correspond to OLS, which are known to produce the best estimates (in a frequentist sense) when the errors are uncorrelated with zero mean and homoscedastic with finite variance. This is the case for the nonoutliers. Our example thus suggests that the choice between the posterior means, medians, MAP or MLE is not crucial for the robust model as well. Second, we observe that in the presence of the 3 outliers (i.e. using the whole sample of size $n = 50$), the results of the LPTN model are barely affected, showing similar results to those excluding the outliers, while the normal model is clearly contaminated by the outliers. This is consistent with our theoretical asymptotic results which indicate agreement with the bulk of the data under the robust model. In particular, the estimate for σ under the LPTN model is about half that arising from the normal model, resulting in much shorter credibility intervals for the robust model. Those patterns in the estimates are typical of the normal and LPTN models. That is reflected in the thorough performance evaluation presented in the next section.

With the posterior in hand, one can take the inference one step further with outlier identification and prediction. The former is first discussed. For each observation $i =$

		Posterior estimates for				
		β_1	β_2	β_3	β_4	σ
Means	LPTN	514.0 [514.5]	1.03 [1.03]	1.12 [1.09]	0.39 [0.36]	47.9 [43.8]
	$\mathcal{N}$	504.9 [514.3]	0.97 [1.02]	1.40 [1.09]	0.70 [0.36]	96.5 [43.1]
Medians	LPTN	514.0 [514.6]	1.03 [1.03]	1.12 [1.09]	0.39 [0.36]	47.4 [43.5]
	$\mathcal{N}$	504.9 [514.3]	0.97 [1.02]	1.40 [1.09]	0.70 [0.36]	95.6 [42.7]
MAP	LPTN	513.0 [513.7]	1.00 [1.01]	1.11 [1.10]	0.40 [0.37]	44.3 [40.8]
	$\mathcal{N}$	504.9 [514.3]	0.97 [1.02]	1.40 [1.09]	0.70 [0.36]	90.1 [40.1]
MLE	LPTN	513.1 [513.8]	1.00 [1.01]	1.11 [1.10]	0.40 [0.37]	44.7 [41.1]
	$\mathcal{N}$	504.9 [514.3]	0.97 [1.02]	1.40 [1.09]	0.70 [0.36]	91.0 [40.5]
CI – LB	LPTN	500.3 [501.9]	0.86 [0.87]	0.81 [0.81]	0.22 [0.21]	36.9 [34.5]
	$\mathcal{N}$	478.1 [501.8]	0.66 [0.87]	0.81 [0.82]	0.38 [0.21]	77.3 [34.4]
CI – UB	LPTN	527.7 [527.0]	1.20 [1.19]	1.42 [1.37]	0.56 [0.52]	59.8 [53.7]
	$\mathcal{N}$	532.2 [526.8]	1.29 [1.18]	2.00 [1.37]	1.02 [0.51]	117.1 [52.7]

Table 2: Posterior means and medians, MAP, MLE and credible intervals (CI – LB and CI – UB), under the LPTN($\rho = 0.95$) and $\mathcal{N}(0, 1)$ assumptions for f; the numbers in square brackets are the estimates based on the 47 nonoutliers only.

$1, \ldots, n$, one can estimate the value fitted by the hyperplane $\mathbf{x}_i^T\boldsymbol{\beta}$, the realisation of the error $y_i - \mathbf{x}_i^T\boldsymbol{\beta}$ and its standardised version $z_i := (y_i - \mathbf{x}_i^T\boldsymbol{\beta})/\sigma$. This can be achieved through their MAP estimates (or MLE) by simply plugging in the MAP estimates (or MLE) of $\boldsymbol{\beta}$ and σ (as given in Table 2) in their expression. Or possibly better, they can be estimated by their posterior mean or median. For this purpose, samples can be directly generated from their posterior distribution through the values of $\boldsymbol{\beta}$ and σ already generated from the RWM algorithm (or obviously, it can be done at the same time the algorithm runs). Consider for instance Home 49, which is valued at $y_{49} = 1{,}000$, the posterior means give fitted values of 704.0 (LPTN) and 759.7 (normal), errors of 296.0 (LPTN) and 240.3 (normal) and standardised errors of 6.28 (LPTN) and 2.52 (normal). We note that the hyperplane is attracted towards the outlier under the normal model, which leads to an estimated error less extreme than that under the LPTN model.

Naturally, large estimates for standardised errors $|z_i|$ indicate strong evidence of outlyingness. A threshold of 2.5 is sometimes recommended to differentiate outliers from nonoutliers, see, e.g., Gervini and Yohai (2002). On this basis, Home 49 appears clearly as an outlier under the LPTN model, while the conclusion is unclear for the normal model.

To provide a measure of outlyingness, we evaluate the probability for a (unrealised) standardised error ϵ_{i_0}/σ — which density is f — to be more extreme than $|z_i|$:

$$\varrho(z_i) := \mathbb{P}(|\epsilon_{i_0}/\sigma| > |z_i|) = \mathbb{P}\left(|\epsilon_{i_0}/\sigma| > |y_i - \mathbf{x}_i^T\boldsymbol{\beta}|/\sigma\right).$$

Under the normal model, we have

$$\varrho^{\mathcal{N}}(z_i) := 2(1 - \Phi(|z_i|)),$$

whereas under the LPTN(ρ) it is

$$\varrho^{\mathrm{LPTN}}(z_i) := \begin{cases} 2(\Phi(\tau) - \Phi(|z_i|)) + 2\varphi(\tau)\tau(\log\tau)\lambda^{-1} & \text{if} \quad |z_i| \leq \tau, \\ 2\varphi(\tau)\tau(\log\tau)\lambda^{-1}\left(\frac{\log\tau}{\log|z_i|}\right)^{\lambda} & \text{if} \quad |z_i| > \tau, \end{cases}$$

where $\tau = 1.96$ and $\lambda = 3.08$ when $\rho = 0.95$, as computed with (3.3).

The measure $\varrho(z_i)$ is a random variable as it is a function of the unknown parameters $\boldsymbol{\beta}$ and σ, and can be estimated *a posteriori* using the same technique as above. In the same spirit as Gervini and Yohai (2002), one can flag observations with estimates for $\varrho(z_i)$ lesser than a chosen threshold. A reasonable threshold, in our opinion, should lie between 0.01 and 0.02. This corresponds to a range of 2.47 to 3.11 of MAP estimates for $|z_i|$ under the LPTN model if ϱ is estimated through its MAP (because this is achieved by plugging in the MAP of $|z_i|$).

If we look again at results of Home 49, the posterior means for $\varrho(z_i)$ give 0.0024 and 0.0208 for the LPTN and normal models, respectively. Home 49 appears again clearly as an outlier under the LPTN model, whereas it is much less convincing for the normal model. At a threshold of 0.02 or less, this observation would not be considered as an outlier. Outlier detection using the wholly robust LPTN model is effective; outliers do not mask each other, a well-known phenomenon arising with nonrobust models typically due to overestimation of the scale σ, and sometimes because of attraction of hyperplanes. The posterior means for the standardised errors z_i are plotted in Figure 5, along with the posterior means for $\varrho(z_i)$ for the three outliers.

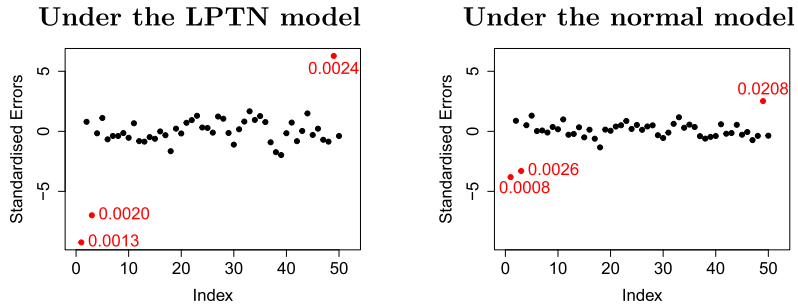

Figure 5: Posterior mean for the standardised errors z_i and outlier identification measures $\varrho(z_i)$, under the LPTN and normal models.

For predicting a future observation, say $Y_{n+1} = \mathbf{x}_{n+1}^T\boldsymbol{\beta} + \epsilon_{n+1}$, we estimate its posterior predictive density by sampling from it through the RWM algorithm as before. For each realisation of $(\boldsymbol{\beta}, \sigma)$ in the Markov chains, we generate ϵ_{n+1} from an LPTN (or a normal for the nonrobust model) centred at 0 with a scale parameter σ, to which we add $\mathbf{x}_{n+1}^T\boldsymbol{\beta}$. We can thus easily compute posterior predictive quantities such as the median, credible intervals, probabilities and so on. Note that the expectation does not exist under the LPTN (because it does not exist for ϵ_{n+1}). MAP can be approximated

from the sample, but because it requires extra work, we suggest using the median for prediction.

If for example we consider the future observation of the typical house with $x_{n+1,2} = x_{n+1,3} = x_{n+1,4} = 0$, the posterior predictive medians for Y_{n+1} are 514.0 and 504.9 under the LPTN and normal models, respectively; they are as expected around the posterior medians of the intercept β_1. The credible intervals are $(417.4, 611.6)$ and $(313.7, 698.6)$ for the LPTN and normal models, respectively. We note the shorter length for the robust model, which is attributable to the robust estimation of the scale σ.

Finally, we easily perform statistical hypothesis testing through Bayes factors. For this, we implement a reversible jump algorithm (Green (1995)) with two models and uniform prior on these. If, for instance, we want to test for hypotheses $H_0 : \beta_4 = 0$ versus $H_1 : \beta_4 \neq 0$, the implementation essentially requires the tuning of an additional RWM algorithm; that for sampling the parameters of the model without x_4. In our example, the Bayes factors are 1.68×10^3 and 1.74×10^3 for the LPTN and normal models, respectively. If we exclude the outliers, they become 2.80×10^3 and 2.12×10^3 for the LPTN and normal models, respectively.

The Bayes factor is a robust measure under the model with a LPTN distribution on the error term. Indeed, Result (a) of Theorem 2.1 states that the marginal $m(\mathbf{y_n})$ behaves like $m(\mathbf{y_k}) \prod_{i=1}^{n} [f(y_i)]^{\ell_i}$. Furthermore, the marginal $m(\mathbf{y_n} \mid H_0)$ behaves like $m(\mathbf{y_k} \mid H_0) \prod_{i=1}^{n} [f(y_i)\ ]^{\ell_i}$, because when the assumptions of Theorem 2.1 are satisfied for the larger model, they are automatically satisfied for the smaller. As a result, the Bayes factor $m(\mathbf{y_n})/m(\mathbf{y_n} \mid H_0)$ behaves like $m(\mathbf{y_k})/m(\mathbf{y_k} \mid H_0)$.

3.4 Performance Evaluation

In this section, we evaluate the performance of the robust LPTN model through a simulation study. We consider the same data set and model as in Section 3.3, but get rid of $\mathbf{y_n}$ which are generated. Several values for ρ are considered: $\rho = 0.80, 0.84, 0.90, 0.93, 0.95$, and 0.98. As in the last Section, it is compared with the nonrobust normal model. We add the Bayesian approach of Box and Tiao (1968) with normal mixtures and the model with the Student distribution. For the latter, we consider different degrees of freedom (df): 1, 2, 4, 6, and 10. We set $\pi(\boldsymbol{\beta}, \sigma) \propto 1$ and estimate the parameters using the MAP, which therefore corresponds to the MLE. The Bayesian methods thus become direct competitors to the frequentist robust estimators like the popular M- and S-estimators. These as well as MM-, REWLSE (the two best frequentist methods according to the recent review by Yu and Yao (2017)) and LTS estimators are included in the simulation study.

The data $\mathbf{y_n}$ are generated through the errors $\epsilon_i \mid \sigma \overset{\mathcal{D}}{\sim} (1/\sigma) f(\epsilon_i/\sigma)$ under the following scenarios:

- **Scenario 0**: $f = \mathcal{N}(0, 1)$,
- **Scenario 1**: $f = 95\%\mathcal{N}(0, 1) + 5\%\mathcal{N}(7, 1)$,
- **Scenario 2**: $f = 90\%\mathcal{N}(0, 1) + 10\%\mathcal{N}(7, 1)$,

- **Scenario 3**: $f = 95\%\,\mathcal{N}(0,1) + 5\%\,\mathcal{N}(3,1)$, where the $\mathbf{x}_i$ of the outliers are modified to make them high-leverage points (the procedure is explained in detail below),
- **Scenario 4**: $f = 90\%\,\mathcal{N}(0,1) + 10\%\,\mathcal{N}(3,1)$, where the $\mathbf{x}_i$ of the outliers are modified to make them high-leverage points.

Nonoutliers are generated from the first mixture component, whereas outliers are generated from the second one. The choice of locations for the outliers aims at producing challenging and interesting situations, where a vast spectrum of behaviours is observed for especially the LPTN and Student models with their different sets of parameters ρ and df. Scenarios 2 and 4 are studied to show how performance varies when the number of outliers is doubled, from 5% to 10% of the sample size. For each scenario, we consider two sample sizes: $n = 50$ and $n = 100$. The case $n = 50$ corresponding to the original $\mathbf{x}_1, \ldots, \mathbf{x}_{50}$, 50 additional observations from the explanatory variables are generated in the same fashion as the original ones for the case $n = 100$.

For Scenarios 3 and 4, when an error is generated from the second mixture component (that generating extreme values), say ϵ_{i_0}, we modify one of the coordinates of the associated $\mathbf{x}_{i_0}$ to make the observation an high-leverage point. More precisely, we randomly choose a covariable number, say $j_0 \in \{2, 3, 4\}$, and set $x_{i_0 j_0} = 1.5 \max_i x_{ij_0}$.

The performance of each model/estimator is evaluated through the *premium versus protection* approach of Anscombe and Guttman (1960). This approach consists in computing the premium to pay for using a robust alternative $\mathcal{R}$ to the normal $\mathcal{N}$ when there are no outliers (Scenario 0), and the protection provided by this alternative when the data sets are contaminated (which is likely in the other scenarios). The premium and protections associated with a robust alternative $\mathcal{R}$ are evaluated through the following:

$$\text{Premium}(\mathcal{R}, \hat{\boldsymbol{\beta}}) := \frac{\mathcal{M}_{\mathcal{R}}(\hat{\boldsymbol{\beta}}) - \mathcal{M}_{\mathcal{N}}(\hat{\boldsymbol{\beta}})}{\mathcal{M}_{\mathcal{N}}(\hat{\boldsymbol{\beta}})},$$

$$\text{Protection}(\mathcal{R}, \hat{\boldsymbol{\beta}} \mid \mathcal{S}) := \frac{\mathcal{M}_{\mathcal{N}}(\hat{\boldsymbol{\beta}} \mid \mathcal{S}) - \mathcal{M}_{\mathcal{R}}(\hat{\boldsymbol{\beta}} \mid \mathcal{S})}{\mathcal{M}_{\mathcal{N}}(\hat{\boldsymbol{\beta}} \mid \mathcal{S})},$$

where $\mathcal{S}$ represents the scenario under which the protection is evaluated (1, 2, 3 or 4), and $\mathcal{M}_{\mathcal{N}}(\hat{\boldsymbol{\beta}} \mid \mathcal{S})$, for instance, denotes an error measure $\mathcal{M}$ for estimating $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ using the normal model $\mathcal{N}$, in Scenario $\mathcal{S}$. The scenario is not specified for the premium because it does not vary; it is Scenario 0. The premiums and protections with respect to $\hat{\sigma}$ — $\text{Premium}(\mathcal{R}, \hat{\sigma})$ and $\text{Protection}(\mathcal{R}, \hat{\sigma} \mid \mathcal{S})$ — have the analogous definitions.

We consider two distinct error measures ($\mathcal{M}_{\mathcal{R}}(\hat{\boldsymbol{\beta}} \mid \mathcal{S})$ and $\mathcal{M}_{\mathcal{R}}(\hat{\sigma} \mid \mathcal{S})$) to highlight the difference between them, and also because there is no natural way of combining them. We propose to define $\mathcal{M}_{\mathcal{R}}(\hat{\boldsymbol{\beta}} \mid \mathcal{S})$ as the square root of the expectation with respect to $\mathbf{Y_n}$ (and therefore the estimates associated with each realisation) of the average squared vertical distances between the estimated and true hyperplanes measured at each observation $\mathbf{x}_i$:

$$\mathcal{M}_{\mathcal{R}}(\hat{\boldsymbol{\beta}} \mid \mathcal{S}) := \left(\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^T \hat{\boldsymbol{\beta}} - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right] \right)^{1/2} = \left(\frac{1}{n} \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right] \right)^{1/2},$$

where $\mathbf{X}$ is the design matrix with rows $\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T$. The expression after the second equality provides us with another interpretation. The measure represents an alternative to $(\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})])^{1/2}$, the square root of the trace of the mean square error (MSE) matrix for $\hat{\boldsymbol{\beta}}$. Given that under the normal model $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$, standardisation is applied to $\hat{\boldsymbol{\beta}}$ in our measure. For $\hat{\sigma}$, we simply use the square root of its MSE: $\mathcal{M}_{\mathcal{R}}(\hat{\sigma} \mid \mathcal{S}) := (\mathbb{E}[(\hat{\sigma} - \sigma)^2])^{1/2}$. Note that the expectations are approximated through the simulation of 250,000 vectors $\mathbf{y_n}$.

The premium and protection for a given robust alternative $\mathcal{R}$ in a given scenario $\mathcal{S}$ are therefore the relative increase and decrease in $\mathcal{M}_{\mathcal{R}}(\cdot \mid \mathcal{S})$ due to the use of the robust alternative instead of the normal (the benchmark model), respectively. For each robust alternative, there are four premiums to compute: one for the measure for $\hat{\boldsymbol{\beta}}$ and one for the measure for $\hat{\sigma}$, in the cases $n = 50$ and $n = 100$. There are sixteen protections to compute given that we also do this for Scenarios 1, 2, 3, and 4. The idea is to graphically present the results by plotting the couples $(\text{Premium}(\mathcal{R}, \hat{\boldsymbol{\beta}}), \text{Protection}(\mathcal{R}, \hat{\boldsymbol{\beta}} \mid \mathcal{S}))$ for all robust alternatives. The results for Scenarios 1 and 2 are shown in Figure 6, and those for Scenarios 3 and 4 in Figure 7.

From this *premium versus protection* perspective, a robust alternative dominates another if its premium is smaller and protection larger. This means that in Figures 6 and 7, we are looking for points in the upper left parts. It is noticed that the robust alternatives are all excellent candidates, except maybe for S-estimator that we choose not to show because of its large premium for $\hat{\boldsymbol{\beta}}$ and its same behaviour as MM-estimator for $\hat{\sigma}$. In particular, the presented robust alternatives all handle high-leverage points.

By looking at Figures 6 and 7, we notice that the LPTN curve (in green) dominates the Student curve (in orange), more remarkably for $\hat{\sigma}$, but also for $\hat{\boldsymbol{\beta}}$. We also notice that the optimal values for ρ for the LPTN are around the nonoutlier percentages, i.e. around 0.95 (the second point starting from the lower left corner) in Scenarios 1 and 3, and around 0.90 (the fourth point starting from the lower left corner) in Scenarios 2 and 4. This justifies our suggestion in Section 3.1 for selecting ρ based on prior knowledge about probable proportions of outliers, if users do not have other preferences. The best LPTN models in all scenarios essentially dominate all the other alternatives with respect to $\hat{\sigma}$. As for $\hat{\boldsymbol{\beta}}$, the performance of these LPTN models is among the best. The mixture model appears better in this case, but often by little. The difference varies depending on the number of outliers and the sample size. For instance, look at the LPTN(0.95) in Scenarios 1 and 3 (and also at the scale of the x axis), and notice how the LPTN(0.98) gets closer to the mixture model in these scenarios when doubling the sample size, which makes this model almost the best. This allows to make an interesting remark: for a given percentage of outliers (and therefore of nonoutliers), a larger sample size translates into enhanced protection, because there are more nonoutliers. This is especially true for LPTN models with ρ close to 1.

4 Conclusion

The goal of this paper, which was to provide a solution that reaches gold standards in terms of *premium versus protection* for all parameters, is now achieved. The foundations

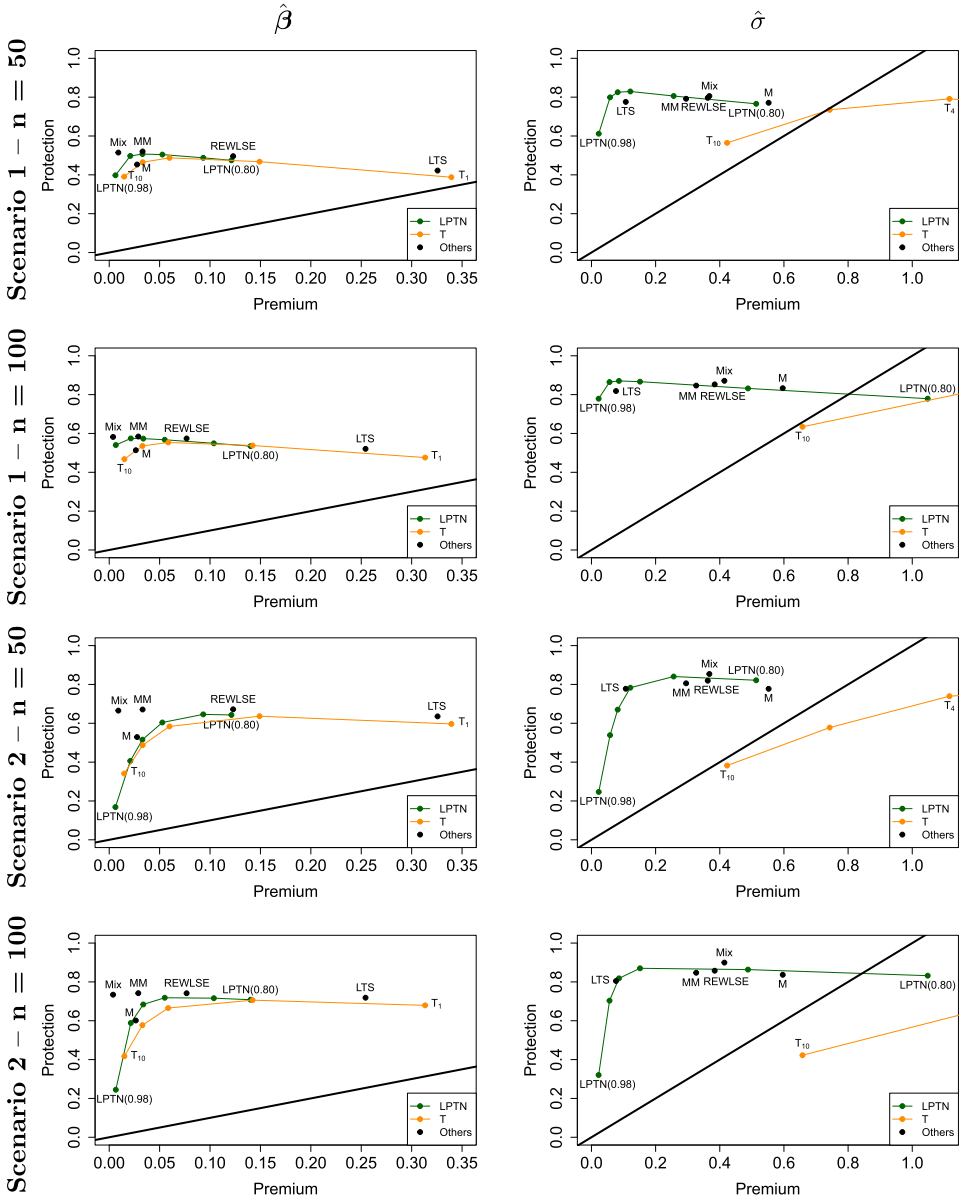

Figure 6: Premiums vs protections in Scenarios 1 and 2, and lines premium = protection to identify the robust alternatives that offer better protections than their premium.

for great protection were established through our main theoretical contribution: the proof of whole robustness results for linear regression. The key result is the convergence of the posterior distribution towards that based on the nonoutliers only when the outliers

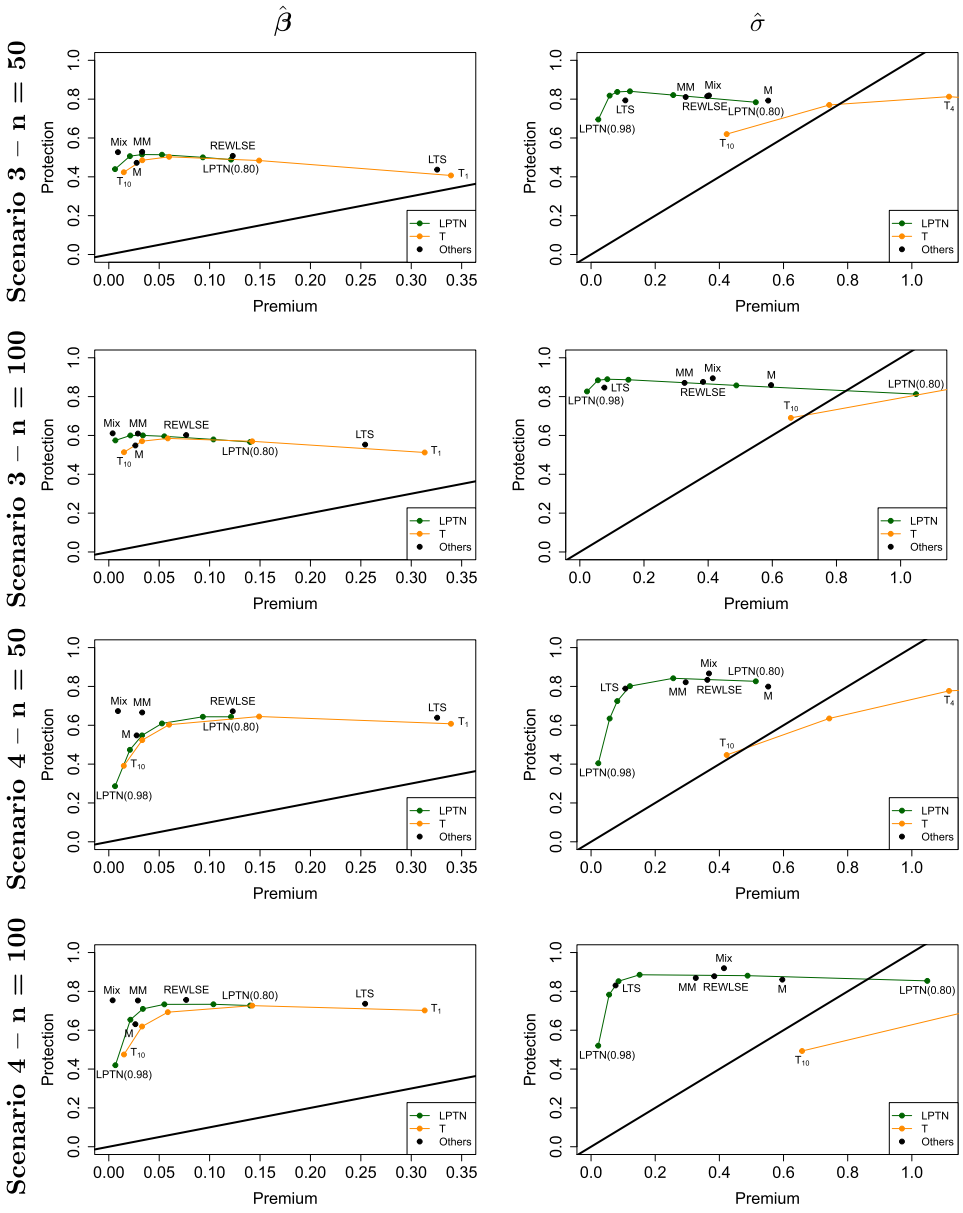

Figure 7: Premiums vs protections in Scenarios 3 and 4, and lines premium = protection to identify the robust alternatives that offer better protections than their premium.

approach plus or minus infinity (Result (c), Theorem 2.1). The robustness results hold under two simple and intuitive conditions. Firstly, the error term must follow a super heavy-tailed distribution, namely a LRVD, to accommodate for the presence of outliers.

Secondly, the number of outliers must not exceed half the sample $n/2$ minus $p - 1/2$ (the number of regression coefficients minus 1/2). This last condition translates into a limiting breakdown point of 0.5 as $n \to \infty$.

Although the whole robustness results are theoretical and asymptotic, their practical relevance has been shown through a comprehensive study of the LPTN model. This specific choice of super heavy-tailed distribution represented our main practical contribution as the resulting model is remarkably efficient and deals with outlying observations in an automatic and sensitive manner, succeeding in achieving low premium in addition to large protection. The procedure for analysing data sets to which it gives rise is also easy to use. These characteristics of the LPTN model make it a particularly appealing Bayesian alternative to the partially robust Student model.

5 Proofs

We in fact provide in this section sketches of the proofs of Proposition 2.1 and Theorem 2.1 for space considerations. The detailed proofs can be found in the supplementary material.

5.1 Proof of Proposition 2.1

Let us pretend for now that the scale parameter is known and that its value is σ_0. To simplify, we denote the posterior density as $\pi(\boldsymbol{\beta} \mid \mathbf{y_n}) := \pi(\boldsymbol{\beta}, \sigma = \sigma_0 \mid \mathbf{y_n})$. To prove that it is proper, we show that the marginal $m(\mathbf{y_n})$ is finite. We have that

$$\begin{aligned}
&\int_{\mathbb{R}^p} \pi(\boldsymbol{\beta}, \sigma_0) \prod_{i=1}^{n} \frac{1}{\sigma_0} f\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma_0}\right) d\boldsymbol{\beta} \\
&\quad \leq B^{n-p+1} \max\left(1, \frac{1}{\sigma_0}\right) \frac{1}{\sigma_0^{n-p}} \int_{\mathbb{R}^p} \prod_{i=1}^{p} \frac{1}{\sigma_0} f\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma_0}\right) d\boldsymbol{\beta} \\
&\quad \leq B^{n-p+1} \max\left(1, \frac{1}{\sigma_0}\right) \frac{1}{\sigma_0^{n-p}} \left| \det \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_p^T \end{pmatrix} \right|^{-1} \prod_{i=1}^{p} \int_{\mathbb{R}} f(u_i)\, du_i,
\end{aligned}$$

using $\pi(\boldsymbol{\beta}, \sigma_0) \leq B \max(1, 1/\sigma_0)$ (by assumption) and $f \leq B$ (because of the assumptions on this PDF), and the changes of variables $u_i = (y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma_0, i = 1, \ldots, p$, B being a positive constant. The last quantity above is finite given that the determinant is different from 0 because all explanatory variables are continuous. Note that this justifies also the assumption mentioned in Remark 2.1 about the full rank of the design matrix when any type of explanatory variables is considered.

An additional integral with respect to σ is added in front when $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n})$ is considered. For σ not too small (bounded from below), it is easy to see that the additional integral is finite because $\max(1, 1/\sigma)$ is bounded and $\sigma^{-(n-p)}$ is integrable if $n - p \geq 2$. This is the case because $n > p+1$ by assumption. For small σ, the proof is more technical

and requires to bound more carefully the densities f than above. See the supplementary material for details.

Proving that $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k})$ is proper is similar. For the moments, we use that

$$\begin{aligned}\mathbb{E}[\sigma^M \mid \mathbf{y_n}] &= \int \sigma^M \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n})\, d\boldsymbol{\beta}\, d\sigma \\ &\leq [m(\mathbf{y_n})]^{-1} B^M \int \pi(\boldsymbol{\beta}, \sigma) \prod_{i=M+1}^{n} \frac{1}{\sigma} f\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right) d\boldsymbol{\beta}\, d\sigma,\end{aligned}$$

using $f \leq B$. This is finite given that $m(\mathbf{y_n}) < \infty$ and the integral is finite because it corresponds to the marginal of $n - M$ observations, and $n - M > p + 1$ by assumption.

For the moments of β_j, it is more technical. Consider the first moment. We would like to compute instead the first moment of $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|$ because $(|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|/\sigma) f(|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|/\sigma) \leq B$ (because of the assumptions on f), and as for the moments of σ, it would be easy to show that the integral is finite. The strategy is to write β_j as $\mathbf{e}_j^T \boldsymbol{\beta}$, where $\mathbf{e}_j$ is a vector of size p having 1 at the j-th position and 0's elsewhere, and to write $\mathbf{e}_j^T$ as a linear combination of p vectors $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}$ to essentially retrieve what we were looking for. See the supplementary material for details.

5.2 Proof of Theorem 2.1

Proof of Result (a). To prove this result, we use that

$$\begin{aligned}\frac{m(\mathbf{y_n})}{m(\mathbf{y_k}) \prod_{i=1}^n [f(y_i)]^{\ell_i}} &= \frac{m(\mathbf{y_n})}{m(\mathbf{y_k}) \prod_{i=1}^n [f(y_i)]^{\ell_i}} \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n})\, d\sigma\, d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \frac{\pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n \left[(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)\right]^{k_i + \ell_i}}{m(\mathbf{y_k}) \prod_{i=1}^n [f(y_i)]^{\ell_i}} d\sigma\, d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k}) \prod_{i=1}^n \left[\frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)}\right]^{\ell_i} d\sigma\, d\boldsymbol{\beta},\end{aligned}$$

and show that this integral converges towards 1 as $\omega \to \infty$. Assuming that we can interchange the limit and the integral, we have that

$$\begin{aligned}&\lim_{\omega \to \infty} \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k}) \prod_{i=1}^n \left[\frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)}\right]^{\ell_i} d\sigma\, d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \lim_{\omega \to \infty} \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k}) \prod_{i=1}^n \left[\frac{(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)}{f(y_i)}\right]^{\ell_i} d\sigma\, d\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \int_0^\infty \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k})\, d\sigma\, d\boldsymbol{\beta} = 1,\end{aligned}$$

using Proposition 2.2 in the second equality, and next Proposition 2.1. Note that the conditions of Proposition 2.1 are satisfied given that $k \geq \ell + 2p - 1 \Rightarrow k \geq p + 2$, assuming that $\ell \geq 1$ (otherwise the proof is trivial) and because $p \geq 2$.

To interchange the limit and the integral, we need to prove that the integrand is bounded by an integrable function of $\boldsymbol{\beta}$ and σ that does not depend on ω. As in Section 5.1, let us set for now the scale parameter to a positive value σ_0. We know that

$$\begin{aligned}&\pi(\boldsymbol{\beta},\sigma_0 \mid \mathbf{y_k})\prod_{i=1}^{n}\left[\frac{(1/\sigma_0)f((y_i-\mathbf{x}_i^T\boldsymbol{\beta})/\sigma_0)}{f(y_i)}\right]^{\ell_i}\\&=[m(\mathbf{y_k})]^{-1}\pi(\boldsymbol{\beta},\sigma_0)\prod_{i=1}^{n}[(1/\sigma_0)f((y_i-\mathbf{x}_i^T\boldsymbol{\beta})/\sigma_0)]^{k_i}\left[\frac{(1/\sigma_0)f((y_i-\mathbf{x}_i^T\boldsymbol{\beta})/\sigma_0)}{f(y_i)}\right]^{\ell_i}.\end{aligned} \tag{5.1}$$

Consider that $\boldsymbol{\beta} \in \mathcal{F}$, a set such that the hyperplanes pass (relatively) close to the nonoutliers (fixed observations), and therefore, (relatively) far to the outliers. In this case, for large enough ω, we have that

$$\prod_{i=1}^{n}\left[\frac{(1/\sigma_0)f((y_i-\mathbf{x}_i^T\boldsymbol{\beta})/\sigma_0)}{f(y_i)}\right]^{\ell_i}$$

is bounded above using Proposition 2.2 because $\mathbf{x}_i^T\boldsymbol{\beta}$ is bounded (recall that $y_i = a_i + b_i\omega$), and the remaining terms on the right-hand side (RHS) in (5.1) give $\pi(\boldsymbol{\beta},\sigma_0 \mid \mathbf{y_k})$ which is integrable.

Consider now that $\boldsymbol{\beta} \in \mathcal{O}$, a set such that the hyperplanes pass (relatively) close to the outliers. The difference is that we are not sure that these hyperplanes do not pass close to the nonoutliers (see Figure 8). In this example, $n = 5$, $k = 4$ and $\ell = 1$, which satisfy the assumptions in Theorem 2.1: $k - \ell = 3 \geq 2(p - 1/2) = 3$. We also have that

$$\frac{(1/\sigma_0)f((y_4-\mathbf{x}_4^T\boldsymbol{\beta})/\sigma_0)}{f(y_5)}$$

is bounded above using again Proposition 2.2 but now because $|y_4 - \mathbf{x}_4^T\boldsymbol{\beta}|$ is close to ω (this is explained in greater detail in the supplementary material). Note that it would not be true if $\mathbf{x}_1 = \mathbf{x}_4$, which is why we require to have enough of different vectors $\mathbf{x}_i$ in Remark 2.2. The remaining terms on the RHS in (5.1) are

$$[m(\mathbf{y_k})]^{-1}\pi(\boldsymbol{\beta},\sigma_0)\prod_{i=1(i\neq 4)}^{n}[(1/\sigma_0)f((y_i-\mathbf{x}_i^T\boldsymbol{\beta})/\sigma_0)],$$

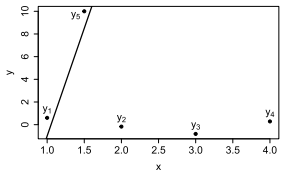

Figure 8: Example of a case where the line passes close to a nonoutlier and an outlier.

which after multiplying and dividing by the right marginal is proportional to the posterior density based on y_1, y_2, y_3, y_5, which is integrable given that $4 = n - \ell \geq p + 2 = 2p + \ell - 1 = 4$. This justifies the assumption on the number of nonoutliers in Theorem 2.1 given by $k = n - \ell \geq 2p + \ell - 1$.

The strategy to do the proof in general is to rewrite the domain of $\boldsymbol{\beta}$ (which is $\mathbb{R}^p$) as a finite number of mutually exclusive sets, in which it is always possible to proceed as above. The function to bound thus becomes a finite sum, where each term is bounded above by integrable function. When σ is free, an additional level of technicalities is added because $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|$ can be large, but not $|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|/\sigma$. See the supplementary material for all the details. □

Proof of Result (b). We have that

$$|\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n}) - \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k})| = \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k}) \left| \frac{m(\mathbf{y_k})}{m(\mathbf{y_n})} \prod_{i=1}^{n} [(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)]^{\ell_i} - 1 \right|.$$

The absolute value on the RHS converges to 0 as $\omega \to \infty$ uniformly on $(\boldsymbol{\beta}, \sigma) \in [-\vartheta, \vartheta]^p \times [1/\eta, \eta]$ using Proposition 2.2 and Result (a), for any $\vartheta \geq 0$ and $\eta \geq 1$. On this set, $\pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k})$ is bounded using the assumptions on the prior and f and the fact that $m(\mathbf{y_k})$ is finite. This concludes the proof. □

Proof of Result (c). Result (c) is a direct consequence of Result (b) using Scheffé's theorem (see Scheffé (1947)). See the supplementary material for details. □

Proof of Result (d). Result (d) is proved through a mix of the strategies used to show Result (a) and that the moments exist in Proposition 2.1. Assuming that we can interchange the limit and the integral, we have

$$\begin{aligned}
\lim_{\omega \to \infty} \mathbb{E}[\sigma^M \mid \mathbf{y_n}] &= \lim_{\omega \to \infty} \int_0^\infty \int_{\mathbb{R}^p} \sigma^M \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n}) \, d\boldsymbol{\beta} \, d\sigma \\
&= \int_0^\infty \int_{\mathbb{R}^p} \lim_{\omega \to \infty} \sigma^M \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n}) \, d\boldsymbol{\beta} \, d\sigma \\
&= \int_0^\infty \int_{\mathbb{R}^p} \sigma^M \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_k}) \, d\boldsymbol{\beta} \, d\sigma = \mathbb{E}[\sigma^M \mid \mathbf{y_k}],
\end{aligned}$$

using Result (b). Again, we have to prove the integrand is bounded by an integrable function of $\boldsymbol{\beta}$ and σ that does not depend on ω. To achieve this, we bound above $\sigma^M \pi(\boldsymbol{\beta}, \sigma \mid \mathbf{y_n})$ by a constant times a function similar to the one that is shown to be bounded by an integrable function of $\boldsymbol{\beta}$ and σ in the proof of Result (a). See the supplementary material for details. We proceed with the same strategy for $\mathbb{E}[\beta_j^M \mid \mathbf{y_n}]$. □

Supplementary Material

A New Bayesian Approach to Robustness Against Outliers in Linear Regression – Supplementary Material (DOI: 10.1214/19-BA1157SUPP; .pdf).

References

Andrade, J. A. A. and O'Hagan, A. (2011). "Bayesian Robustness Modelling of Location and Scale Parameters." *Scandinavian Journal of Statistics*, 38(4): 691–711. MR2859745. doi: https://doi.org/10.1111/j.1467-9469.2011.00750.x. 390

Anscombe, F. J. and Guttman, I. (1960). "Rejection of Outliers." *Technometrics*, 2(2): 123–147. MR0123405. doi: https://doi.org/10.2307/1266540. 405

Box, G. E. P. and Tiao, G. C. (1968). "A Bayesian Approach to Some Outlier Problems." *Biometrika*, 55(1): 119–129. MR0225427. doi: https://doi.org/10.1093/biomet/55.1.119. 389, 390, 404

Bunke, O., Milhaud, X., et al. (1998). "Asymptotic behavior of Bayes estimates under possibly incorrect models." *Annals of Statistics*, 26(2): 617–644. MR1626075. doi: https://doi.org/10.1214/aos/1028144851. 399

Desgagné, A. (2013). "Full Robustness in Bayesian Modelling of a Scale Parameter." *Bayesian Analysis*, 8(1): 187–220. MR3036259. doi: https://doi.org/10.1214/13-BA808. 394

Desgagné, A. (2015). "Robustness to Outliers in Location–Scale Parameter Model using Log-Regularly Varying Distributions." *Annals of Statistics*, 43(4): 1568–1595. MR3357871. doi: https://doi.org/10.1214/15-AOS1316. 390, 391, 394, 395, 398

Desgagné, A. and Gagnon, P. (2019). "Bayesian robustness to outliers in linear regression and ratio estimation." *Brazilian Journal of Probability and Statistics*, 33(2): 205–221. ArXiv:1612.05307. MR3919021. doi: https://doi.org/10.1214/17-bjps385. 391

Gagnon, P., Desgagné, A., and Bédard, M. (2019). "A New Bayesian Approach to Robustness Against Outliers in Linear Regression – Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1157SUPP. 399

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741. 390

Gervini, D. and Yohai, V. J. (2002). "A class of robust and fully efficient regression estimators." *Annals of Statistics*, 30(2): 583–616. MR1902900. doi: https://doi.org/10.1214/aos/1021379866. 390, 402, 403

Green, P. J. (1995). "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika*, 82(4): 711–732. MR1380810. doi: https://doi.org/10.1093/biomet/82.4.711. 404

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." emphBiometrika, 57(1): 97–109. MR3363437. doi: https://doi.org/10.1093/biomet/57.1.97. 401

Huber, P. J. (1973). "Robust Regression: Asymptotics, Conjectures and Monte Carlo." *Annals of Statistics*, 799–821. MR0356373. 390

Karamata, J. (1930). "Sur un mode de croissance régulière des fonctions." *Mathematica (Cluj)*, 4: 38–53. 394

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of State Calculations by Fast Computing Machines." *Journal of Chemical Physics*, 21: 1087. 401

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics." *Handbook of Markov Chain Monte Carlo*, 2(11). MR2858447. 401

O'Hagan, A. and Pericchi, L. (2012). "Bayesian heavy-tailed models and conflict resolution: A review." *Brazilian Journal of Probability and Statistics*, 26(4): 372–401. MR2949085. doi: https://doi.org/10.1214/11-BJPS164. 389

Peña, D., Zamar, R., and Yan, G. (2009). "Bayesian likelihood robustness in linear models." *Journal of Statistical Planning and Inference*, 139(7): 2196–2207. MR2507981. doi: https://doi.org/10.1016/j.jspi.2008.10.012. 390

Rousseeuw, P. J. (1985). "Multivariate estimation with high breakdown point." *Mathematical Statistics and Applications*, 37(8): 283–297. MR0851060. 390

Rousseeuw, P. J. and Yohai, V. J. (1984). "Robust regression by means of S-estimators." In *Robust and Nonlinear Time Series Analysis*, 256–272. Springer. MR0786313. doi: https://doi.org/10.1007/978-1-4615-7821-5_15. 390

Scheffé, H. (1947). "A Useful Convergence Theorem for Probability Distributions." *Annals of Mathematical Statistics*, 434–438. MR0021585. doi: https://doi.org/10.1214/aoms/1177730390. 412

West, M. (1984). "Outlier Models and Prior Distributions in Bayesian Linear Regression." *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 46(3): 431–439. MR0790630. 390, 391

Yohai, V. J. (1987). "High breakdown-point and high efficiency robust estimates for regression." *Annals of Statistics*, 15: 642–656. MR0888431. doi: https://doi.org/10.1214/aos/1176350366. 390

Yu, C. and Yao, W. (2017). "Robust linear regression: A review and comparison." *Communications in Statistics – Simulation and Computation*, 46(8): 6261–6282. MR3740779. doi: https://doi.org/10.1080/03610918.2016.1202271. 404

Acknowledgments

The authors acknowledge support from NSERC (Natural Sciences and Engineering Research Council of Canada), FRQNT (Le Fonds de recherche du Québec – Nature et technologies) and SOA (Society of Actuaries). They also acknowledge enlightening discussions with Professor Judith Rousseau about consistency of Bayes estimates. They finally thank an anonymous referee and an associate editor for their helpful comments.