Bayesian Analysis (2023) **18**, Number 3, pp. 841–864

Robustness Against Conflicting Prior Information in Regression[*]

Philippe Gagnon[†]

Abstract. Including prior information about model parameters is a fundamental step of any Bayesian statistical analysis. It is viewed positively by some as it allows, among others, to quantitatively incorporate expert opinion about model parameters. It is viewed negatively by others because it sets the stage for subjectivity in statistical analysis. Certainly, it creates problems when the inference is skewed due to a conflict with the data collected. According to the theory of conflict resolution (O'Hagan and Pericchi, 2012), a solution to such problems is to diminish the impact of conflicting prior information, yielding inference consistent with the data. This is typically achieved by using heavy-tailed priors. We study both theoretically and numerically the efficacy of such a solution in a regression framework where the prior information about the coefficients takes the form of a product of density functions with known location and scale parameters. We study functions with regularly-varying tails (Student distributions), log-regularly-varying tails (as introduced in Desgagné (2015)), and propose functions with slower tail decays that allow to resolve any conflict that can happen under that regression framework, contrarily to the two previous types of functions. The code to reproduce all numerical experiments is available online.[‡]

Keywords: Bayesian statistics, built-in robustness, constant-tailed priors, heavy-tailed distributions, weak convergence, whole robustness.

1 Introduction

1.1 Context

In Bayesian analysis, prior information about the parameters of a regression model is included using prior distributions. Consider a model $Y \sim \mathbb{P}_{\eta,\boldsymbol{\psi}}$, with $\eta := \mathbf{x}^T\boldsymbol{\beta}$ being a linear predictor. For this regression model, the parameters are $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$, where $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ are the regression coefficients, with p a positive integer, and $\boldsymbol{\psi}$ is a vector formed of, e.g., scale or shape parameters; $\mathbf{x}$ is a known vector of covariates. This regression framework encompasses linear regression, generalized linear models (GLMs) and generalized additive models (when estimated using a spline representation). In this paper, we study the impact on statistical inference of prior information in conflict with the data collected for different types of prior distributions. Our study rests heavily on the form of the prior distributions which will be seen to be a product form where each

[*] The author acknowledges support from NSERC (Natural Sciences and Engineering Research Council of Canada) and FRQNT (Le Fonds de recherche du Québec – Nature et technologies).

[†] Department of Mathematics and Statistics, Université de Montréal, Canada, philippe.gagnon.3@umontreal.ca

[‡] See ancillary files on arXiv:2110.09556.

https://doi.org/10.1214/22-BA1330

regression-coefficient density has known location and scale parameters, justifying the introduction of such a study within a regression framework. We focus on situations where the prior information that is in conflict is about regression coefficients, the latter being typically of main interest which makes them more likely to be assigned informative prior distributions. Additionally, we focus on situations where the prior distributions on the coefficients are used to include prior information about the latter, not to regularize the model (contrarily to in, e.g., Johnstone and Silverman (2004), Park and Casella (2008), and Carvalho et al. (2010)), even though the study conducted here may be helpful to develop regularization strategies. Furthermore, we focus on linear regression as a special case of the general regression framework described above. This will allow to state precise theoretical results about the behaviour of the posterior distribution in conflicting situations, depending on the type of prior distributions employed. We will explain why and how the results presented apply in the general regression framework.

From now on, we thus consider that $Y = \mathbf{x}^T\boldsymbol{\beta}+\sigma\varepsilon$ with $\varepsilon \sim f$, which is equivalent to $Y \sim (1/\sigma)f((\,\cdot\, - \mathbf{x}^T\boldsymbol{\beta})/\sigma)$, where ε is a standardized error term, $\sigma > 0$ is a scale parameter and f is a distribution; to simplify, f is also used to denote the probability density function (PDF) associated to the distribution. When all covariates are continuous (i.e. when they all take values in uncountable totally-ordered sets), it is recommended to define the prior distribution of the regression coefficients using a conditional-independence structure (see, e.g., West (1984) and Raftery et al. (1997)):

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\psi}) = \pi(\boldsymbol{\beta} \mid \sigma) := \prod_{j=1}^{p} \pi_j(\beta_j \mid \sigma) := \prod_{j=1}^{p} \frac{\lambda_j}{\sigma} g_j\left(\frac{\lambda_j}{\sigma}(\beta_j - \mu_j)\right), \tag{1}$$

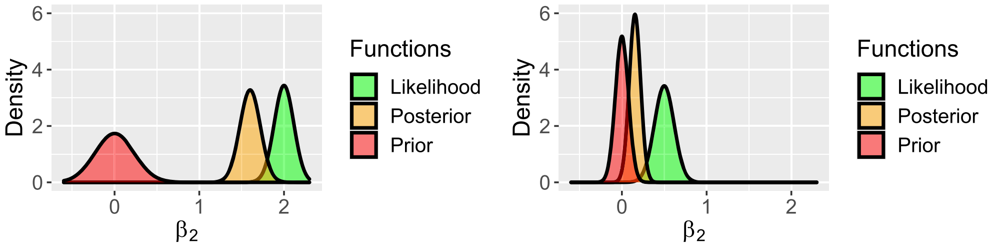
where all g_j are strictly positive bounded density functions that are symmetric with respect to 0, and $\mu_j \in \mathbb{R}$ and $\sigma/\lambda_j > 0$ play the role of location and scale parameters, respectively; μ_j and λ_j are considered to be known and chosen by the user. In the following, we consider to simplify that all covariates are continuous; the theoretical results hold even when this is not the case, but under more technical assumptions. Note that, to simplify the notation, g_j is also used to denote the distribution associated to the density.

Determining the outcome of conflicting prior information under a general structure of dependence in between the coefficients requires a multivariate analysis and depends strongly on the structure of dependence. The conditional-independence structure presented above allows to simplify the problem and transform the multivariate analysis into several univariate analyses, in addition to enabling the exploitation of existing conflict-resolution techniques that are based on univariate heavy-tailed distributions (and for which the relevant literature will be presented when describing the techniques below). A general and multivariate analysis to determine the outcome of conflicting prior information is beyond the scope of this manuscript.

1.2 Conflicts

In normal linear regression, where $f = \mathcal{N}(0,1)$, conjugate priors are often employed, i.e. $g_j = \mathcal{N}(0,1)$ in (1) and σ^2 follows an inverse-gamma distribution. A prior is in conflict with the likelihood when the areas where these function have high densities are significantly different (Figure 1). When both the prior and the likelihood are normal (given σ), an undesirable compromise follows: the posterior concentrates its mass on an area in between those with high prior and likelihood densities. This is a consequence of the slimness of the normal tails: the area where the likelihood function has high density is in the tails of the prior density which have an exponential decay, penalizing extremely for such parameter values, and the same holds if we inverse the role of the prior and likelihood in the previous statement. The areas with high prior and likelihood densities thus become *a posteriori* less probable than an area in between, representing how a conflict is dealt with by that Bayesian modelling and an ineffective way of resolving a conflict. Indeed, the posterior distribution is not consistent with either of the sources of information, be it the prior or the likelihood. Here we consider that the data model is well specified and that the data can be trusted; the information about the parameters carried by the data is thus favoured to the prior information when they conflict. Therefore, we consider that a conflict is (effectively) resolved when the conflicting prior information is discarded so as to yield a posterior distribution consistent with the data (Figure 2).

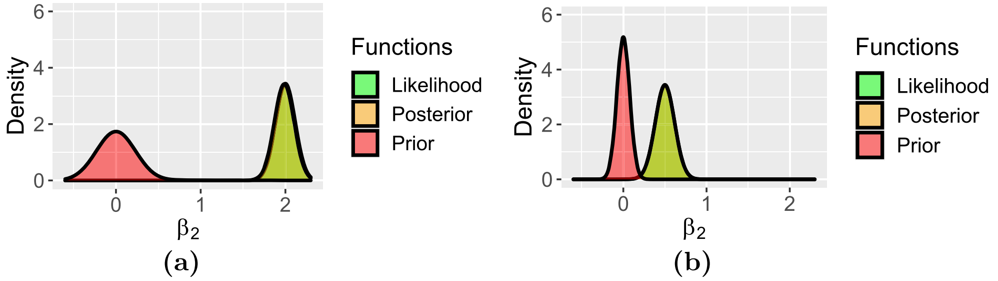
We acknowledge that the assumption of a well specified data model and that the data can be trusted is strong, but we make this assumption in order to be able to focus on robustness against conflicting prior information. We can, for instance, allow for some sort of misspecification and a potential presence of extreme/erroneous data on top of conflicting prior information by considering that the data set may contain outliers, and obtain similar theoretical results as those presented in the next sections. This is because we allow for the regression model to have an heavy-tailed error distribution. It is however beyond the scope of this manuscript to analyse the situation of potential presence of outliers and present related results.



(a) **(b)**

Figure 1: Two examples of conflicts where $y = 0 + \beta_2 x_2 + \sigma\varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$, $g_2 = \mathcal{N}(0,1)$, the sample size is $n = 100$, the prior on σ is an inverse-gamma with shape and scale parameters of $n/2$ each, the variables are standardized, and: (a) $\mu_2 = 0$, $\lambda_2 = \sqrt{n}/2$ and $\hat{\beta}_2^{\mathrm{OLS}} = 2$, (b) $\mu_2 = 0$, $\lambda_2 = 1.5\sqrt{n}$ and $\hat{\beta}_2^{\mathrm{OLS}} = 0.5$; $\hat{\beta}_2^{\mathrm{OLS}}$ is the ordinary-least-squares (OLS) estimate which corresponds to the maximum likelihood estimate in that case; in this figure, the likelihood function is normalized to make it a PDF.

Plots (a) and (b) in Figures 1 and 2 are meant to represent two distinct conflicting situations: (a) one where the conflict is due to a prior location that is significantly different than that of the likelihood, and (b) one where it is due to a extremely small prior scaling. We analyse both situations theoretically and numerically in the next sections. The theoretical analysis will be conducted under an asymptotic regime. In the



(a) **(b)**

Figure 2: Same two examples of conflicts as in Figure 1 with the only difference being that in: (a) $g_2 =$ LPTN as defined in Section 2 with $\rho = 0.95$; (b) $g_2 =$ CTN as defined in Section 2 with $\varrho = 0.98$.

first situation, the asymptotic regime corresponds to one where the distance between the red and green areas in Figure 1 (a) increases without bounds, which is mathematically modelled by the red one moving away, i.e. $\mu_j \to \pm\infty$; in the second situation, we will consider that $\lambda_j \to \infty$. It will be seen that a prior distribution which does lead to a resolution of conflict in the first situation does not necessarily in the second one. The first situation can be think of as one where a practitioner was wrong about the parameter location, but incorporated a moderate confidence by using a moderate prior scaling. In the second situation, in addition to being wrong about the parameter location (but much less severely than in the first situation), the practitioner was also overly confident; this conflicting situation could have been avoided by using a less concentrated prior. The latter is also true in the first situation, but the prior would need to be much less concentrated to compensate for the significant difference in location. That is why in one case we consider that the problematic aspect is the location, whereas we consider that it is the scaling in the other one.

A natural way to achieve effective conflict resolution is to have recourse to heavy-tailed distributions; in situations like those presented in Figure 2, the areas where the likelihood functions have high densities are still in the tails of the prior densities, but more weight is assigned to those tails, thus penalizing less for such extreme situations. This strategy dates back to de Finetti (1961) with a first analysis in Lindley (1968), followed by an introduction of a formal theory in Dawid (1973), Hill (1974) and O'Hagan (1979). For a recent review of Bayesian heavy-tailed models and conflict resolution, see O'Hagan and Pericchi (2012). In the latter paper, it is noted that there exists a gap between the models formally covered by the theory of conflict resolution and models commonly used in practice. The latest developments focus on situations where the conflicting information is carried by outlying data points in location-scale models (Desgagné, 2015) and linear regression (Desgagné and Gagnon, 2019; Gagnon et al., 2020, 2021; Hamura et al., 2022; Gagnon and Hayashi, 2022). The present paper contributes to the expansion of the theory of conflict resolution by covering conflicting prior information in a regression framework.

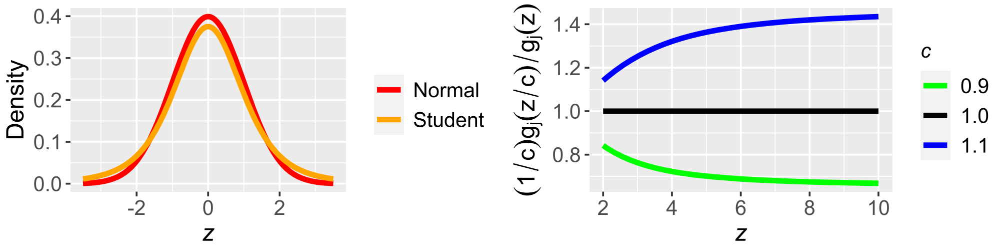
We consider that in the ideal situation where it is guaranteed that the priors will not conflict with the data that will be collected, prior information about the regression coefficients is included by setting $g_j = \mathcal{N}(0,1)$, which is the favoured choice in practice. Given that we consider that the data model is well specified and that the data can be trusted, the distributions that we alter to achieve effective conflict resolution are thus the g_j's. A desideratum of the resulting heavy-tailed priors is to yield similar inference to the informative light-tailed priors they replace in the absence of conflict. In the following, we study three alternatives to the normal distribution with three different types of tail decays: a first one with regularly-varying tails, a second one with log-regularly-varying tails (Desgagné, 2015), and a third one with constant tails. They are all presented in Section 2 in which an overview of their advantages and disadvantages is also provided. In Section 3, their efficacy is precisely characterized through theoretical results. An extensive simulation study is next provided in Section 4 to show how these theoretical results translate in practice. The manuscript finishes in Section 5 with retrospective comments. All proofs of theoretical results are deferred to Appendix A (supplementary material (Gagnon, 2022)). Some details of the simulation study are presented in Appendix B (supplementary material).

2 Heavy-tailed priors

We start in Section 2.1 by presenting the main characteristics of the most commonly employed alternative to the normal distribution in conflict resolution, the Student distribution. Even in the least problematic situation, which is that where the conflict is due to a prior location that is significantly different than that of the likelihood, it will be seen to partially resolve conflicts. We next provide a description of the *log-Pareto-tailed normal* (LPTN) distribution in Section 2.2, which has the ability to wholly discard the prior information in that situation. This distribution was introduced by Desgagné (2015). Its density exactly matches that of the standard normal on the interval $[-\tau, \tau]$, where $\mathbb{P}(-\tau \leq \mathcal{N}(0,1) \leq \tau) = \rho$. Outside of this area, the tails of this continuous density are log-regularly varying (Desgagné, 2015), and behave as log-Pareto tails, i.e. $(1/|z|)(1/\log|z|)^\theta$, hence its name. The only free parameter of this distribution is ρ: the parameter θ is a function of ρ and τ, the latter being itself a function of ρ. Even with such heavy tails, the LPTN distribution leads to an ineffective conflict resolution when the conflict is due to a small prior scaling. In response to this problem, we introduce in Section 2.3 the *constant-tailed normal* (CTN) distribution, which, like the LPTN distribution, has a density that matches that of the standard normal on a central interval, but with constant tails.

2.1 Student distribution

The Student distribution is without a doubt the favourite heavy-tailed alternative to the standard normal distribution. A reason for this is because its density shares important characteristics with the standard normal one, like a bell shape and symmetry around 0. We show this in Figure 3 (a) for a Student distribution with 4 degrees of freedom, which represents a good compromise between heavy tails and close similarity with the normal distribution. In Figure 3 (b), we show how the ratio $(1/c)g_j(z/c)/g_j(z)$ behaves as $z \to \infty$ when g_j is the PDF of a Student distribution with 4 degrees of freedom to graphically illustrate its regularly-varying property, a property that is discussed in greater detail below.



(a) **(b)**

Figure 3: (a) PDFs of the standard normal distribution and the Student distribution with 4 degrees of freedom, (b) ratios of two Student PDFs with 4 degrees of freedom where the PDF at the numerator as an additional scale parameter of value c.

Employing Student prior distributions instead of normal ones for conflict resolution in regression has been explored before; see, e.g., West (1984) and Mutlu et al. (2019). However, the focus of previous papers was different than that of the current one, which is to compare that alternative to normal prior distributions to other alternatives through an extensive theoretical and numerical analysis. In West (1984), for instance, the focus is rather to study the use of the Student distribution in a context of robustness against outliers; the Student is viewed as a member of a specific family of alternatives to normal distributions, that of scale mixtures of normal distributions.

The tails of the Student density are regularly varying, implying that for any fixed $\lambda_j, \sigma, \beta_j$,

$$\lim_{\mu_j \to \pm\infty} \frac{\frac{\lambda_j}{\sigma} g_j\left(\frac{\lambda_j}{\sigma}(\beta_j - \mu_j)\right)}{g_j(\mu_j)} = \lim_{\mu_j \to \pm\infty} \frac{\lambda_j}{\sigma} \left(\frac{\gamma + \mu_j^2}{\gamma + (\lambda_j/\sigma)^2(\beta_j - \mu_j)^2}\right)^{\frac{\gamma+1}{2}} = \left(\frac{\sigma}{\lambda_j}\right)^{\gamma}, \tag{2}$$

where γ is the degrees of freedom. Examining the limiting behaviour of prior densities is an important step in understanding the limiting behaviour of the posterior distribution in conflicting situations. Indeed, given that the posterior density is the normalized product of the prior densities and the likelihood function, the limit above suggests that a conflicting prior density (due to a significantly different location) behaves in the limiting posterior distribution like $(\sigma/\lambda_j)^\gamma g_j(\mu_j) \propto \sigma^\gamma$. The theoretical results in Section 3 precisely characterize the behaviour of the limiting posterior distribution, depending on the conflicting situation and the priors employed. With a Student prior distribution, conflicting information is partially rejected as a trace remains, σ^γ. Ideally, conflicting information is wholly rejected as its source becomes increasingly remote (West, 1984), which translates into a prior density which behaves asymptotically like $g_j(\mu_j) \propto 1$. This explains why we say that the Student distribution only partially resolves conflicts due to significantly different locations. The existence of that trace is a consequence of employing a prior density with insufficiently heavy tails. Indeed, it will be seen in Section 2.2

that the limit of the ratio in (2) when instead setting g_j to a LPTN distribution is 1. The trace has an impact on the limiting posterior variability of all coefficients which is seen to be more or less significant depending on the degrees of freedom, the sample size and the number of conflicting prior densities (this is shown explicitly in Section 4). When the sample size is large relatively to the degrees of freedom and the number of conflicting prior densities, as in the numerical experiment of Section 4, the impact is small. The insufficiently heavy tails however make the convergence to the limiting posterior distribution slower, comparatively with other alternatives with heavier tails, implying a slower partial resolution of conflicts. We finish this section by noting that the Student prior density converges to a point mass at μ_j when $\lambda_j \to \infty$, which makes it ineffective at resolving conflicts due to extremely small prior scalings.

2.2 LPTN distribution

The density of the LPTN distribution is as follows:

$$g_{\text{LPTN}}(z) := \begin{cases} \varphi(z) & \text{if} \quad |z| \leq \tau, \\ \varphi(\tau)\, \frac{\tau}{|z|} \left(\frac{\log \tau}{\log |z|}\right)^{\theta} & \text{if} \quad |z| > \tau, \end{cases}$$

where $z \in \mathbb{R}$, and $\tau > 1$ and $\theta > 1$ are functions of a parameter $\rho \in (2\Phi(1) - 1, 1) \approx (0.6827, 1)$ with

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

φ, Φ and Φ^{-1} being the PDF, cumulative distribution function (CDF) and inverse CDF of a standard normal, respectively. A LPTN density with $\rho = 0.95$ is presented in Figure 4 (a). This choice of value for ρ yields, like the Student with 4 degrees of freedom in Figure 3 (a), a good compromise between heavy tails and close similarity with the normal distribution. In Figure 4 (b), we show how the ratio $(1/c)g_j(z/c)/g_j(z)$ behaves as $z \to \infty$ when g_j is the PDF of a LPTN distribution with $\rho = 0.95$ to graphically illustrate its log-regularly-varying property, a property that is discussed in greater detail below.

As was seen in Figure 2 (a), this slightly modified version of the normal distribution with $\rho = 0.95$ can resolve conflicts very effectively; the likelihood function and posterior density are indeed on top of each other in that figure. The parameter ρ, chosen by the user, represents the mass of the central part that exactly matches the $\mathcal{N}(0,1)$ density. The value 0.95 has been seen to be a good choice for robustness against outliers in linear regression (see Gagnon et al. (2020) and Gagnon et al. (2021)). We analyse the impact of the value of ρ in the present context in Section 4.

An advantage of Student distributions over LPTN distributions is that their densities are smooth. Indeed, we see in Figure 4 (a) that in order to obtain a density that exactly matches that of the standard normal on an interval, while having heavier tails and being continuous, the LPTN density has to decrease quicker than the normal one

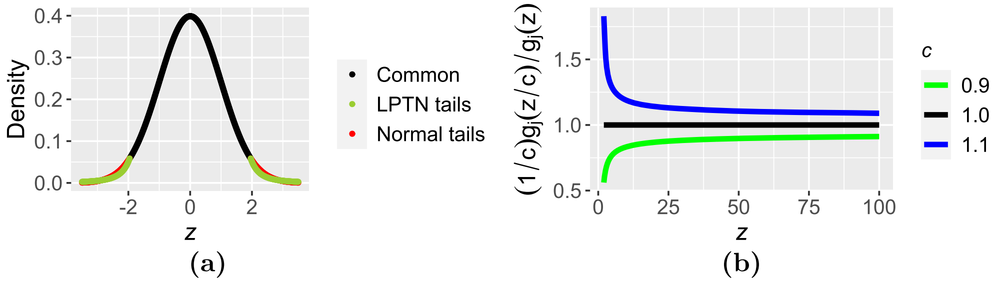


Figure 4: (a) PDFs of the standard normal distribution and the LPTN distribution with $\rho = 0.95$, (b) ratios of two LPTN PDFs with $\rho = 0.95$ where the PDF at the numerator as an additional scale parameter of value c.

for a short interval beyond $|z| = \tau$, making the derivative of the LPTN density discontinuous at $|z| = \tau$. The smoothness of a posterior density has an impact on the efficiency of the numerical methods used to approximate integrals with respect to the associated posterior distribution. With densities having discontinuous derivatives, one might wonder if it is even possible to apply numerical methods explicitly exploiting gradients of log posterior densities, like Metropolis-adjusted Langevin algorithms (Roberts and Tweedie, 1996) and Hamiltonian Monte Carlo (HMC, Duane et al. (1987)), which both are Markov-chain Monte Carlo methods. Given that the discontinuity points of the LPTN derivative have null measure, these methods can be applied. HMC has in fact been employed to sample from the resulting posterior distributions to compute estimates and posterior variances in Section 4 and no problems have been encountered.

The main advantage of LPTN distributions over Student distributions is that the limit of the ratio of densities analogous to (2) is equal to 1, as established in the next proposition, showing their ability at effectively resolving conflicts due to significantly different locations.

Proposition 1 (Asymptotic location-scale invariance)**.** *If $g_j = g_{LPTN}$, we have that for any fixed $\lambda_j, \sigma, \beta_j$,*

$$\lim_{\mu_j \to \pm\infty} \frac{\frac{\lambda_j}{\sigma} g_j\left(\frac{\lambda_j}{\sigma}(\beta_j - \mu_j)\right)}{g_j(\mu_j)} = \lim_{\mu_j \to \pm\infty} \frac{|\mu_j|}{|\beta_j - \mu_j|}\left(\frac{\log|\mu_j|}{\log(\lambda_j/\sigma)|\beta_j - \mu_j|}\right)^{\theta} = 1.$$

The property of asymptotic location-scale invariance is shared by all log-regularly-varying distributions (LRVDs, Desgagné (2015)). Most members of this family of distributions supported on the real line are distributions whose density tails were not originally log-Pareto ones (like the standard normal distribution), but their tails were replaced to reach the desired tail decay, i.e. $(1/|z|)(1/\log|z|)^{\theta}$ (like the LPTN distribution). The strategy of replacing the tails of a light-tailed prior density by heavy ones to attain an asymptotic location-scale invariance can thus be applied even when the light-tailed prior is not a normal. A distribution which is LRVD, but with originally log-Pareto tails, is the log transformation of a Pareto distribution. This distribution is not often employed because its density has a spike at zero, and is thus less appealing than smooth bell curves like normal and Student densities.

Another advantage of LPTN distributions over Student distributions is that its density is even more similar to the normal one, yielding more similar inferences in the absence of conflict, as will be seen in Section 4. Although there are advantages in using LPTN prior distributions, there are also disadvantages; one has been mentioned above, but the main disadvantage is that they do not allow, like all LRVDs, to resolve conflicts due to large λ_j. Indeed, for any fixed $\beta_j, \mu_j \in \mathbb{R}$ and $\sigma > 0$, with $\beta_j \neq \mu_j$, if $g_j = g_{\text{LPTN}}$ and λ_j is large enough,

$$\begin{aligned} \frac{\lambda_j}{\sigma} g_j\left(\frac{\lambda_j}{\sigma}(\beta_j - \mu_j)\right) &= \varphi(\tau)\,\frac{\lambda_j}{\sigma}\,\frac{\tau}{(\lambda_j/\sigma)|\beta_j - \mu_j|}\left(\frac{\log\tau}{\log[(\lambda_j/\sigma)|\beta_j - \mu_j|]}\right)^{\theta} \\ &= \varphi(\tau)\,\frac{\tau}{|\beta_j - \mu_j|}\left(\frac{\log\tau}{\log\lambda_j}\right)^{\theta}\left(\frac{1}{1 + [\log|\beta_j - \mu_j|/\sigma]/\log\lambda_j}\right)^{\theta}, \quad (3) \end{aligned}$$

which is asymptotically equivalent as $\lambda_j \to \infty$ to

$$\varphi(\tau)\,\frac{\tau}{|\beta_j - \mu_j|}\left(\frac{\log\tau}{\log\lambda_j}\right)^{\theta} \propto \frac{1}{|\beta_j - \mu_j|}.$$

Analogously to Student priors in the previous section (but not for the same type of conflict), conflicting information is partially rejected as a trace remains, $|\beta_j - \mu_j|^{-1}$. The latter includes information about the location significantly differently than the normal distribution does. Additionally, the rightmost term in (3) converges to 1 slowly (because the speed at which $[\log|\beta_j - \mu_j|/\sigma]/\log\lambda_j$ vanishes is dictated by that at which $\log\lambda_j$ goes to infinity). This implies that the conflicting information is slowly (in addition to partially) rejected, which will be observed empirically in Section 4. For all these reasons, we consider that LPTN prior distributions are ineffective at resolving conflicts due to small prior scalings, motivating the introduction of different heavy-tailed alternatives to normal prior distributions.

2.3 CTN distribution

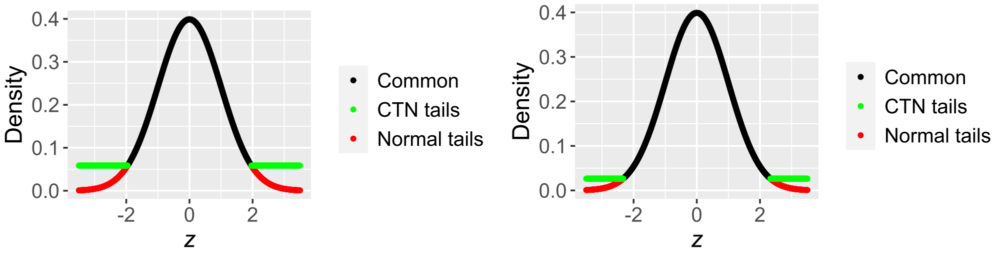
The distribution that is introduced to resolve a conflict due to either a prior location significantly different than that of the likelihood or small prior scalings is the CTN distribution. Its density is as follows:

$$g_{\text{CTN}}(z) := \begin{cases} \varphi(z) & \text{if} \quad |z| \leq \kappa, \\ \varphi(\kappa) & \text{if} \quad |z| > \kappa, \end{cases} \quad (4)$$

where $z \in \mathbb{R}$ and κ is a function of the sole free parameter of the CTN distribution, $\varrho \in (0, 1)$, with an analogous definition to τ in the previous section:

$$\kappa = \Phi^{-1}((1 + \varrho)/2) = \{\kappa : \mathbb{P}(-\kappa \leq Z \leq \kappa) = \varrho \text{ for } Z \sim \mathcal{N}(0, 1)\}.$$

A CTN density with $\varrho = 0.95$ is presented in Figure 5 (a). We observe in Figure 5 (a) that even though the CTN density with $\varrho = 0.95$ matches the standard normal one on the same interval as the LPTN with $\rho = 0.95$ (Figure 4 (a)), its level of similarity with the standard normal density is much lower. Increasing the value of ϱ from 0.95 to, for instance, 0.98 alleviates this issue, as seen in Figure 5 (b), at the price of a slower conflict resolution (but not significantly slower as will be seen in Section 4). The effectiveness of CTN priors with $\varrho = 0.98$ at resolving conflicts due to small prior scalings was shown in Figure 2 (b).



(a) **(b)**

Figure 5: PDFs of the standard normal distribution and the CTN distribution with (a) $\varrho = 0.95$ and (b) $\varrho = 0.98$.

The main disadvantage of CTN prior distributions is that they are improper; they thus cannot be used when there are more parameters than observations. Another disadvantage is that the derivative of their densities is discontinuous, like that of LPTN densities. The discontinuity points of the CTN derivative have null measure, like those of the LPTN derivative, implying that samplers like HMC can be employed. The estimates and posterior variances needed for Section 4 have been computed using HMC, and no problems have been encountered, like with the posterior distributions resulting from LPTN priors.

The main advantage of CTN prior distributions is their limiting behaviour: for any fixed β_j and σ

$$\frac{\lambda_j}{\sigma} g_j\left(\frac{\lambda_j}{\sigma}(\beta_j - \mu_j)\right) \to \frac{\lambda_j}{\sigma}\varphi(\kappa) \propto \frac{1}{\sigma},$$

whenever: i) $\mu_j \to \pm\infty$ and λ_j is fixed, or ii) $\lambda_j \to \infty$ and μ_j is fixed but different than β_j. The conflict resolution is not perfect as the term $1/\sigma$ does not disappear in the limiting posterior density. This term comes from the form of the prior (with a scale parameter given by σ/λ_j), and is not a consequence of insufficiently heavy tails as was the case for Student prior distributions for conflicts due to significantly different locations and LPTN prior distributions for conflicts due to small prior scalings. It should be seen as a flaw of the method. The tails of CTN densities are indeed sufficiently heavy, and allow to resolve any conflict that can happen under our regression framework. A consequence of their sufficiently heavy tails is that they yield a fast convergence towards the limiting posterior distribution, as will be seen in Section 4.

As mentioned for Student prior distributions which yield a similar trace (recall (2)), the remaining term $1/\sigma$ for CTN prior distributions has an impact on the limiting posterior variability of all coefficients which is seen to be more or less significant depending on the sample size and the number of conflicting prior densities. However, contrarily to Student prior distributions, the impact does not increase with the level of similarity between CTN prior distributions and normal ones; recall that the trace left asymptotically by a conflicting Student prior distribution (due to significantly different locations) is σ^γ and that the level of similarity with a normal prior is controlled through γ. The level of similarity between CTN prior distributions and normal ones is controlled through κ, and its value does not have an impact on the trace left asymptotically by a conflicting CTN prior distribution; the trace is $1/\sigma$ regardless of the value of κ.

Note that, in an ideal setting where one knows how many conflicting prior distributions there are, one can multiply the prior of σ (that would ideally be used in a situation where there is no conflict) by σ with a power corresponding to the number of conflicting priors to perfectly resolve the conflict. In practice, one may have prior beliefs about that number, but cannot be sure about it. Consequently, we recommend to not alter the prior of σ as it can cause more harm than good.

3 Theoretical results

In this section, we present three theoretical results. For the presentation of these results, it is required to introduce a proper mathematical framework and details about the model. This is done in Section 3.1 and the results follow. In Section 3.2, we consider an ideal situation where one has access to full information about the conflict, namely, which prior distributions are in conflict and why. This situation is unrealistic but it allows to show what is an ideal conflict resolution in a regression framework. Next, in Section 3.3, we present a result in a situation where one has access to partial information, namely, that there is no conflict due to small prior scalings. This is a more realistic scenario that can be think of as one where a practitioner includes information about the regression coefficients, but the practitioner is cautious while doing it, in the sense that the practitioner uses moderate to large prior scalings. In practice (as we saw in Figure 1 (a)), there is no certainty that there will be no conflict due to a prior location significantly different than that of the likelihood, even when using moderate to large prior scalings, and we consider that the practitioner wants to be protected against this risk. The last situation is the most common one where a practitioner wants to include information about the regression coefficients and thus sets values for all μ_j and λ_j. While having no reason to believe *a priori* that a conflict will occur (and thus while having no information regarding a potential conflict), the practitioner wants to be protected. A result in that situation is presented in Section 3.4.

Throughout the current section, we aim to characterize with theoretical results how conflicts are dealt with asymptotically when using heavy-tailed priors. Theoretical results like those in Bunke and Milhaud (1998) allow to study the limiting behaviour of posterior distributions resulting from heavy-tailed priors under another asymptotic regime than that study here, namely the large-sample regime $n \to \infty$. Even if some heavy-tailed priors presented in Section 2 are non-smooth, it can be proved that they yield posterior distributions that concentrate around the correct parameter values as $n \to \infty$ and posterior estimates that are consistent and asymptotically normal, provided that the priors are non-conflicting (μ_j and λ_j are all held fixed) and the data model is regular enough (which is the case, for instance, for linear regression and GLMs). This means that if there is no conflicting prior information, whether heavy-tailed priors are used or not does not have an impact asymptotically as $n \to \infty$ on the posterior distributions and estimates.

3.1 Mathematical framework

We first precisely describe the asymptotic regime under which the theoretical results in the next subsections are stated. We assume that possibly some $\mu_j \to \pm\infty$ and/or some $\lambda_s \to \infty$ (with s different than j). To analyse separately the effect of misspecified locations and scalings and to simplify the analysis, we indeed consider that when $\mu_j \to \pm\infty$, λ_j is fixed, and when $\lambda_s \to \infty$, μ_s is fixed. We more precisely consider that for all j,

- $\mu_j = a_j + b_j\omega$, with $a_j, b_j \in \mathbb{R}$,
- $\lambda_j = c_j + d_j\omega$ with $c_j > 0$ and $d_j \geq 0$,

under the constraint that $b_j \neq 0$ for conflicting locations, but 0 otherwise, and $d_j > 0$ for conflicting scalings, but 0 otherwise, with $b_j = 0$ if $d_j > 0$ and $d_j = 0$ if $b_j \neq 0$, and we let $\omega \to \infty$. This framework allows, for instance, for conflicting scalings to decrease (because $\lambda_j \to \infty$) at different speeds, meaning that it represents situations where there may be several conflicting scalings, but their values, while being extreme, are not the same.

We now present the model assumptions and introduce required notation. Consider that we observed n data points from a dependent variable, denoted by $y_1, \ldots, y_n \in \mathbb{R}$, where n is a positive integer. Consider also that we have access to n vectors of $p \in \{2, 3, \ldots\}$ covariates, denoted by $\mathbf{x}_1 := (x_{11}, \ldots, x_{1p})^T, \ldots, \mathbf{x}_n := (x_{n1}, \ldots, x_{np})^T \in \mathbb{R}^p$, where in particular $x_{11} = \ldots = x_{n1} = 1$ to introduce an intercept in the model. As typically done in linear regression, we treat these vectors as known constants, i.e. not as realizations of random variables, contrarily to $y_1, \ldots, y_n$. The posterior distribution is thus conditional on the latter only.

In linear regression, the random variables Y_i are modelled as $Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma\varepsilon_i$, $i = 1, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_n \in \mathbb{R}$ are random standardized errors. We assume that the $n + 2$ random variables $\varepsilon_1, \ldots, \varepsilon_n$, $\boldsymbol{\beta}$ and σ are independent, implying that

$$\varepsilon_i \mid \boldsymbol{\beta}, \sigma \overset{\mathcal{D}}{=} \varepsilon_i \sim f, \quad i = 1, \ldots, n,$$

where "$\overset{\mathcal{D}}{=}$" denotes an equality in distribution. This latter assumption is common.

The resulting posterior density is given by

$$\pi_\omega(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) := \pi_\omega(\boldsymbol{\beta}, \sigma) \left[\prod_{i=1}^{n} (1/\sigma) f((y_i - \mathbf{x}_i^T\boldsymbol{\beta})/\sigma)\right] \Bigg/ m_\omega(\mathbf{y}), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0, \quad (5)$$

where $\mathbf{y} := (y_1, \ldots, y_n)^T$, $\pi_\omega(\,\cdot\,,\cdot\,)$ is the prior density and

$$m_\omega(\mathbf{y}) := \int_{\mathbb{R}^p} \int_0^\infty \pi_\omega(\boldsymbol{\beta}, \sigma) \left[\prod_{i=1}^n (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma) \right] \mathrm{d}\sigma \, \mathrm{d}\boldsymbol{\beta}.$$

A dependence on ω (implying a potential presence of conflict) is highlighted using a subscript. The definition of the posterior distribution in (5) only makes sense when the density is integrable, and thus the marginal density $m_\omega(\mathbf{y})$, playing the role of a normalizing constant in this case, is finite. We provide in the next subsections sufficient conditions ensuring that this is the case for all ω and for the limiting posterior density. The limiting posterior distribution is denoted by $\overline{\pi}(\,\cdot\,,\cdot \mid \mathbf{y})$ and its normalizing constant is $\overline{m}$. Their expressions depend on the situations presented in the next subsections.

We now present regularity conditions on f. We assume that:

- f is a strictly positive continuous PDF that is symmetric with respect to 0;
- all parameters of f, if any, are known;
- there exists a threshold above which the function ξ defined by $z \mapsto zf(z)$ is monotonic;
- there exists a positive constant M such that $f/g_{\text{LPTN}} \leq M$.

Examples of PDFs satisfying these conditions include those of normal, Laplace, Student (with pre-specified degrees of freedom) and LPTN (with pre-specified ρ) distributions. The last assumption above on f is about the tail decay of f; it must be at most as slow as that of g_{LPTN}. This implies that our results are also valid when heavy-tailed error distributions are used for robustness against outliers.

The assumptions on $\pi_\omega(\,\cdot \mid \sigma)$ have been presented in Section 1.1. Denote by $\pi(\,\cdot\,)$ the prior of σ that would ideally be used in a situation where there is no conflict. The assumptions on this density depend on the situations presented in the next subsections and will thus be stated in these subsections.

We finish this section by defining the index set of conflicting priors: $\mathrm{C} := \{j : b_j \neq 0 \text{ or } d_j > 0\}$. The index set of non-conflicting priors is thus given by: C^c. We also define two subsets of C: $\mathrm{C}_\mathrm{b} := \{j : b_j \neq 0\}$ and $\mathrm{C}_\mathrm{d} := \{j : d_j > 0\}$, which are such that $\mathrm{C}_\mathrm{b} \bigcup \mathrm{C}_\mathrm{d} = \mathrm{C}$ and $\mathrm{C}_\mathrm{b} \bigcap \mathrm{C}_\mathrm{d} = \varnothing$.

3.2 Full information

Consider that we have set values for all μ_j and λ_j, and that we are provided with the set C. We use the latter to set all g_j accordingly. More precisely, for all $j \in \mathrm{C}_\mathrm{b}$, we set $g_j = g_{\text{LPTN}}$, and for all $j \in \mathrm{C}_\mathrm{d}$, we set $g_j = g_{\text{CTN}}$. We consider that non-conflicting priors, with $j \in \mathrm{C}^\mathsf{c}$, are set to proper distributions with densities having tails not more heavy than those of LPTN densities. Given that we are provided with the set C and we set some priors to CTN distributions (if $\mathrm{C}_\mathrm{d} \neq \varnothing$), we adjust the prior on σ to get rid of the trace left asymptotically by CTN distributions, i.e. the resulting prior density is proportional to $\pi(\sigma)$ multiplied by $\sigma^{|\mathrm{C}_\mathrm{d}|}$. We assume that $\pi(\sigma)$ is bounded above by a constant or a constant times $1/\sigma$, for all $\sigma > 0$, which allows for most proper prior distributions and improper prior densities proportional to $1/\sigma$ or 1.

Theorem 1. *Assume that for all $j \in C_b$, $g_j = g_{LPTN}$, for all $j \in C_d$, $g_j = g_{CTN}$, and for all $j \in C^{\mathsf{c}}$, the positive constant M can be chosen such that $g_j/g_{LPTN} \leq M$. Assume that the prior on σ has a density that is proportional to $\sigma^{|C_d|}\pi(\sigma)$, for all $\sigma > 0$. Assume that the constant M can be chosen such that $\pi(\sigma) \leq \max(M, \sigma^{-1}M)$. Assume that $n + |C^{\mathsf{c}}| \geq 2p - 1 + |C_b|$. Under the framework described in Section 3.1 (recall in particular the form of the prior distribution (1), the definition of the posterior distribution (5), and that $\mu_j = a_j + b_j\omega$ and $\lambda_j = c_j + d_j\omega$), and as $\omega \to \infty$, the posterior distribution converges:*

$$\pi_\omega(\,\cdot\,,\cdot \mid \mathbf{y}) \to \overline{\pi}(\,\cdot\,,\cdot \mid \mathbf{y}),$$

where

$$\overline{\pi}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) := \pi(\sigma) \prod_{j \in C^{\mathsf{c}}} \pi_j(\beta_j \mid \sigma) \left[\prod_{i=1}^{n} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)\right] \Big/ \overline{m}(\mathbf{y}),$$

$$\boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0, \qquad (6)$$

with

$$\overline{m}(\mathbf{y}) = \int_{\mathbb{R}^p} \int_0^\infty \pi(\sigma) \prod_{j \in C^{\mathsf{c}}} \pi_j(\beta_j \mid \sigma) \left[\prod_{i=1}^{n} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)\right] \mathrm{d}\sigma \, \mathrm{d}\boldsymbol{\beta}.$$

The result of Theorem 1 essentially follows from a characterization of the asymptotic behaviour of the marginal distribution:

$$\frac{m_\omega(\mathbf{y})}{\prod_{j \in \mathrm{C_b}} g_j(\mu_j) \prod_{j \in \mathrm{C_d}} \lambda_j \, g_j(\kappa)} \to \overline{m}(\mathbf{y}), \qquad (7)$$

with $m_\omega(\mathbf{y})/[\prod_{j \in \mathrm{C_b}} g_j(\mu_j) \prod_{j \in \mathrm{C_d}} \lambda_j \, g_j(\kappa)] < \infty$ and $\overline{m}(\mathbf{y}) < \infty$ (implying that the posterior distributions are proper); recall that $\kappa = \Phi^{-1}((1 + \varrho)/2)$, where ϱ is the parameter of the CTN distribution. From the characterization in (7) we can, indeed, prove that the posterior density converges pointwise, which in turn allows to prove the convergence of the posterior distribution using Scheffé's theorem (see Scheffé (1947)).

To prove (7), we exploit the proof of Theorem 2.1 in Gagnon et al. (2020). That paper is about robustness to outliers in linear regression. Theorem 2.1 in Gagnon et al. (2020) characterizes the limiting behaviour of the posterior distribution as some $y_i \to \pm\infty$. The prior on all parameters is assumed to be non-conflicting with a joint prior density bounded above by $\max(M, \sigma^{-1}M)$. To exploit the proof of that result in Gagnon et al. (2020), we write

$$\frac{m_\omega(\mathbf{y})}{\prod_{j \in \mathrm{C_b}} g_j(\mu_j) \prod_{j \in \mathrm{C_d}} \lambda_j \, g_j(\kappa) \, \overline{m}(\mathbf{y})}$$

as an integral that is seen to converge to 1 if we are allowed to interchange the limit $\omega \to \infty$ and the integral. We verify that we are allowed to do this by using Lebesgue's dominated convergence theorem. The problem then becomes to prove that the integrand is bounded by an integrable function of $\boldsymbol{\beta}$ and σ that does not depend on ω. This is the main difficulty. We show that it is sufficient to bound above

$$\overline{\pi}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) \left[\prod_{j \in \mathrm{C_b}} \frac{\pi_{j,\omega}(\beta_j \mid \sigma)}{g_j(\mu_j)}\right] = \overline{\pi}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) \left[\prod_{j \in \mathrm{C_b}} \frac{\frac{\lambda_j}{\sigma} g_j\left(\frac{\lambda_j}{\sigma}(\beta_j - \mu_j)\right)}{g_j(\mu_j)}\right] \tag{8}$$

by an integrable function of $\boldsymbol{\beta}$ and σ that does not depend on ω.

To fit within the framework of Gagnon et al. (2020), it suffices to treat $\lambda_j \mu_j$ for $j \in \mathrm{C^c} \bigcup \mathrm{C_b}$ as an observation from the dependent variable and λ_j that multiplies β_j for j in the same set as a vector of covariates where the other covariates are all equal to 0. Then we realize that a technical and lengthy part of the proof of Theorem 2.1 in Gagnon et al. (2020) is devoted to a proof that a function of which (8) is a special case is bounded by an integrable function of $\boldsymbol{\beta}$ and σ that does not depend on ω. The main challenge is that the terms $g_j(\mu_j)$ in the denominator of the product in (8) goes to 0 as $\omega \to \infty$; the strategy is thus to find a way to get rid of these terms by finding an upper bound for any $(\boldsymbol{\beta}, \sigma)$.

When β_j is far from μ_j, we can use Proposition 1 to bound $\pi_{j,\omega}(\beta_j \mid \sigma)/g_j(\mu_j)$ in (8). But this does not work when β_j is not far from μ_j. In this case, we have to use a density $(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)$ in $\overline{\pi}(\boldsymbol{\beta}, \sigma \mid \mathbf{y})$ which is presumably close to 0 when β_j is not far from $\mu_j \to \pm\infty$, and bound above $(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)/g_j(\mu_j)$. Note that we can also use a prior density with $j \in \mathrm{C^c}$. The proof is based on a decomposition of the parameter space into disjoint sets; for each of these sets, we are able to identify in which case we precisely are. In the case where β_j is not far from μ_j, it is shown that the associated hyperplanes pass close to at most $p - 1$ non-conflicting sources (data points in the case of Gagnon et al. (2020)) using that $\mathbf{x}_i$ can be written as a linear combination of p other covariate vectors and the explicit form of the linear-regression model. Other non-conflicting data points are thus such that $(1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)$ are close to 0. The argument is technical and essentially consists in isolating cases where the parameters are such that the densities of *conflicting* sources are evaluated in the tails and those where the parameters are instead such that the densities of *non-conflicting* sources are evaluated in the tails; there is no reason to believe that the result does not hold for other regression models, and in particular, in the general regression framework presented in Section 1.1 including GLMs, perhaps under different assumptions. We believe that even though it turns out that the assumptions are indeed different, they will be similar in essence.

The assumption that $n + |\mathrm{C^c}| \geq 2p - 1 + |\mathrm{C_b}|$ is essentially to ensure that the non-conflicting sources of information are dominant. It is a consequence of: when β_j is not far from μ_j, possibly $|\mathrm{C_b}|$ non-conflicting sources are required to get rid of terms $g_j(\mu_j)^{-1}$ in (8), and imagine that the number of non-conflicting sources left is $2p - 1$, then $p - 1$ of them may be close to hyperplanes such that β_j is not far from μ_j and, using the decomposition in Gagnon et al. (2020), it is shown that p non-conflicting sources are sufficient to obtain an integrable function.

By looking at (6), we see that we get rid asymptotically of all the conflicting priors and no trace is left; the resulting limiting posterior distribution is that with improper Jeffreys priors $\pi_j(\beta_j \mid \sigma) \propto 1$, for $j \in \mathrm{C}$. We thus have a characterization of the limiting behaviour of the posterior distribution/density and estimates like maximum a posteriori probability (MAP) estimates and posterior medians. It is possible to show under additional mild assumptions that the posterior expectations and the joint posterior distribution of a model indicator and parameters in a context of variable selection converge as well. All these results thus characterize the limiting behaviour of a variety of Bayes estimators. Analogous results hold in the situations that are presented in Sections 3.3 and 3.4.

3.3 Partial information

Now, consider that we have set values for all μ_j and λ_j, and we know that $\mathrm{C_d} = \varnothing$. In practice, the situation is rather that a practitioner is confident that there will be no conflict due to small scalings. We now describe how to set the priors in this case and the limiting behaviour of the posterior distribution if the practitioner turns out to be right. Given that each of the priors on the regression coefficients is exposed to a potential conflict due to a prior location significantly different than that of the likelihood, we set $g_j = g_{\mathrm{LPTN}}$ for all j. The advantage here is that, because no CTN distribution is used, no adjustment on the prior of σ is required to yield, as in the previous section, a limiting posterior distribution without a trace of conflict and with improper Jeffreys priors $\pi_j(\beta_j \mid \sigma) \propto 1$, for $j \in \mathrm{C}$. The prior on σ is thus set to $\pi(\,\cdot\,)$ and we assume that $\pi(\sigma)$ is bounded above by a constant or a constant times $1/\sigma$, for all $\sigma > 0$, as before.

Theorem 2. *Assume that $C_d = \varnothing$. Assume that for all j, $g_j = g_{LPTN}$. Assume that the prior on σ is $\pi(\,\cdot\,)$ and that $\pi(\sigma) \leq \max(M, \sigma^{-1}M)$ for all $\sigma > 0$. Assume that $n+|C^c| \geq 2p-1+|C_b|$. Under the framework described in Section 3.1 (recall in particular the form of the prior distribution (1), the definition of the posterior distribution (5), and that $\mu_j = a_j + b_j\omega$ and $\lambda_j = c_j + d_j\omega$), and as $\omega \to \infty$, the posterior distribution converges:*

$$\pi_\omega(\,\cdot\,,\cdot \mid \mathbf{y}) \to \overline{\pi}(\,\cdot\,,\cdot \mid \mathbf{y}),$$

where $\overline{\pi}(\,\cdot\,,\cdot \mid \mathbf{y})$ is defined as in (6).

Theorem 2 is an adaptation of Theorem 1 in which it is considered that $\mathrm{C_d} = \varnothing$, which implies that $\mathrm{C_b} = \mathrm{C}$. Also, the proof of Theorem 2 is an adaptation of that of Theorem 1. For the same reasons as those explained in Section 3.2, we thus believe that Theorem 2 holds in the general regression framework presented in Section 1.1 including GLMs, perhaps under different, yet similar, assumptions. A difference between Theorem 2 and Theorem 1 is that, because we do not know which of the priors will be in conflict (if any) and thus set all $g_j = g_{\mathrm{LPTN}}$, the prior distributions in the limiting posterior for $j \in \mathrm{C^c}$ are thus all LPTN distributions; in Theorem 1, they can be selected to be otherwise, provided that they are proper distributions with densities having tails not more heavy than those of LPTN densities. Using LPTN prior distributions is perhaps not the first choice for a practitioner, but this comes with protection, as seen in Theorem 2.

It is possible to prove a similar result to Theorem 2 if instead we set g_j to a Student distribution for each j. The difference is that the limiting posterior distribution is defined otherwise than in (6). It is instead such that

$$\overline{\pi}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) \propto \sigma^{|\mathrm{C_b}|\gamma} \pi(\sigma) \prod_{j \in \mathrm{C^c}} \pi_j(\beta_j \mid \sigma) \left[\prod_{i=1}^{n} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)\right],$$

reflecting that Student distributions asymptotically leave a trace in case of conflict, namely σ^γ for each of the conflicting priors, and thus that they only partially resolve conflicts due to significantly different locations.

3.4 No information

In the last scenario, we consider that after setting all μ_j and λ_j, we have no reason to believe that these choices of locations and scalings will create conflicts, but we want to be protected in case it happens. We thus set $g_j = g_{\mathrm{CTN}}$ for all j to be prepared for all eventualities. As mentioned, the main disadvantage of using CTN priors is that they are improper. Their densities do not integrate and each g_j is multiplied by λ_j/σ to yield $\pi_{j,\omega}(\beta_j \mid \sigma)$ (recall (1)). This implies that these densities cannot be used to integrate over $\boldsymbol{\beta}$ when verifying, for instance, that $\pi_\omega(\,\cdot\,, \cdot \mid \mathbf{y})$ is proper; the best that can be done is to bound them by a constant (that possibly depends on ω) that is multiplied by σ^{-p}, and to use the (conditional) densities of $Y_1, \ldots, Y_n$ to integrate over $\boldsymbol{\beta}$ (requiring $n \geq p$). Therefore, in order to obtain a proper posterior distribution, the prior on σ needs to be such that $\int \sigma^{-p}\, \pi(\sigma)\, \mathrm{d}\sigma < \infty$. The good news is that setting $\pi(\,\cdot\,)$ such that σ^2 has an inverse-gamma distribution, as often done in practice (West, 1984; Raftery et al., 1997), implies that $\int \sigma^{-p}\, \pi(\sigma)\, \mathrm{d}\sigma < \infty$ for any p, and choice of shape and scale parameters for the inverse-gamma distribution.

Theorem 3. *Assume that for all j, $g_j = g_{CTN}$. Assume that the prior on σ is $\pi(\,\cdot\,)$ and that it is such that $\int \sigma^{-p}\, \pi(\sigma)\, \mathrm{d}\sigma < \infty$. Assume that $n \geq p$. Under the framework described in Section 3.1 (recall in particular the form of the prior distribution* (1)*, the definition of the posterior distribution* (5)*, and that $\mu_j = a_j + b_j\omega$ and $\lambda_j = c_j + d_j\omega$), and as $\omega \to \infty$, the posterior distribution converges:*

$$\pi_\omega(\,\cdot\,, \cdot \mid \mathbf{y}) \to \overline{\pi}(\,\cdot\,, \cdot \mid \mathbf{y}),$$

where

$$\overline{\pi}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) := \sigma^{-|C|}\, \pi(\sigma) \prod_{j \in C^c} \pi_j(\beta_j \mid \sigma) \left[\prod_{i=1}^{n} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)\right] \Big/ \overline{m}(\mathbf{y}), \qquad (9)$$

$\boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0$, *with*

$$\overline{m}(\mathbf{y}) = \int_{\mathbb{R}^p} \int_0^\infty \sigma^{-|C|}\, \pi(\sigma) \prod_{j \in C^c} \pi_j(\beta_j \mid \sigma) \left[\prod_{i=1}^{n} (1/\sigma) f((y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma)\right] \mathrm{d}\sigma\, \mathrm{d}\boldsymbol{\beta}.$$

The proof of Theorem 3 is much simpler than those of Theorems 1 and 2. While still using Lebesgue's dominated convergence theorem, the term that is sufficient to bound by an integrable function of $\boldsymbol{\beta}$ and σ that does not depend on ω is

$$\overline{\pi}(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) \left[\prod_{j \in \mathrm{C}} \frac{g_{\mathrm{CTN}}\left(\frac{\lambda_j}{\sigma}(\beta_j - \mu_j)\right)}{g_{\mathrm{CTN}}(\kappa)} \right] \leq \overline{\pi}(\boldsymbol{\beta}, \sigma \mid \mathbf{y})\, M^{|\mathrm{C}|},$$

which is thus bounded by a function that does not depend on ω, contrarily to (8). The core of the proof of Theorem 3 is essentially devoted to proving that $\overline{\pi}(\,\cdot\,, \cdot \mid \mathbf{y})$ is proper (requiring $n \geq p$ and that $\int \sigma^{-p}\, \pi(\sigma)\, \mathrm{d}\sigma < \infty$ in our case, as seen in Theorem 3). In other words, Theorem 3 holds for any of the regression model fitting in the general regression framework presented in Section 1.1 including GLMs, provided that the prior distribution exhibit a conditional-independence structure as in (1) and the resulting limiting posterior distribution (9) is proper. With Theorem 3, it is thus even clearer than with Theorems 1 and 2 that the result holds in the general regression framework presented in Section 1.1, perhaps under different, yet similar, assumptions.

As mentioned previously, a weakness of using CTN prior distributions is that a trace asymptotically remains in case of conflict, namely $\sigma^{-|\mathrm{C}|}$, as seen in (9). This is similar to what happens when using Student prior distributions (recall the discussion at the end of Section 3.3). However, there is an important difference: CTN prior distributions are effective against all types of conflicting situations, including those due to conflicting prior scalings, contrarily to Student prior distributions. Also, the degree of discrepancy between the resulting limiting posterior distribution and the ideal one (obtained in Theorems 1 and 2), measured through the exponent of σ, does not depend on the level of similarity between the CTN distributions used and the standard normal (measure through the parameter ϱ). With Student prior distributions, the degree of discrepancy between the resulting limiting posterior distribution and the ideal one depends on the degrees of freedom γ. Recall that we recommend to not alter the prior of σ with the aim of correcting for a discrepancy because we do not know $|\mathrm{C}|$ *a priori* and thus an adjustment can cause more harm than good.

4 Simulation study

A goal with this section is to show the impact of using an informative prior instead of a non-informative one, especially in the situation where the former is conflicting. Another goal is to identify suitable values for the hyperparameters of the heavy-tailed priors. We achieve all that through a simulation study; it suggests that $\gamma = 4$ degrees of freedom for Student prior distributions, $\rho = 0.95$ for LPTN prior distributions and $\varrho = 0.98$ for CTN prior distributions are suitable values. For the simulation study, we consider the normal linear regression model, i.e. $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sigma \varepsilon_i$ with $f = \mathcal{N}(0, 1)$. For the reasons mentioned in Section 3, we expect the results to be similar for other regression models, such as GLMs. To simplify, we consider that the covariates are orthogonal and that the variables are standardized, i.e. $(1/n)\sum_{i=1}^n y_i = 0$ and $(1/n)\sum_{i=1}^n y_i^2 = 1$, $(1/n)\sum_{i=1}^n x_{ij} = 0$ and $(1/n)\sum_{i=1}^n x_{ij}^2 = 1$ for all j (except for $j = 1$ for which $(1/n)\sum_{i=1}^n x_{i1} = 1$), and

$\sum_{i=1}^{n} x_{ij}x_{is} = 0$ for $j \neq s$. Under this framework, the likelihood function exhibits a hierarchical and product form and is proportional to:

$$\frac{1}{\sigma^{n-p}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2\right) \prod_{j=1}^{p} \frac{1}{\sigma} \exp\left(-\frac{n}{2\sigma^2}(\beta_j - \hat{\beta}_j)^2\right),$$

where $\| \cdot \|_2$ is the Euclidean norm and $\hat{\mathbf{y}} := \mathbf{X}\hat{\boldsymbol{\beta}}$, $\mathbf{X}$ being the design matrix and $\hat{\boldsymbol{\beta}} := (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ the OLS estimate, which in this case is such that $\hat{\beta}_j = (1/n)\sum_{i=1}^{n} x_{ij}y_i$.

With this likelihood form, setting any prior $\pi_{j,\omega}(\,\cdot \mid \sigma)$ on β_j with $\mu_j = \hat{\beta}_j$ yields the same marginal posterior distributions of the other coefficients regardless of the values of $\hat{\beta}_j$ and λ_j, as long as $\hat{\mathbf{y}}$ is the same. To simplify, we consider that $\mu_j = \hat{\beta}_j$ for all coefficients except one, namely β_2, that will be used to show the impact of different choices for μ_2, λ_2 and g_2 to achieve our aforementioned goals. We also consider to simplify that $\hat{\boldsymbol{\beta}} = \mathbf{0}$, so that the marginal posterior distribution of β_2 only depends on n (not on p and the covariate data points); we set $n = 100$. It can be readily verified that $n > 3$ is sufficient to ensure a proper posterior distribution, even if the prior distributions of β_2 and σ are improper Jeffreys priors. This condition is satisfied in the simulation study, and thus to simplify, we set the prior on σ to the Jeffreys prior: $\pi(\sigma) \propto 1/\sigma$. The non-informative Jeffreys prior on β_2 will serve as a benchmark, i.e. $\pi_{2,\omega}(\beta_2 \mid \sigma) \propto 1$, implying a posterior mean and variance of 0 and $1/(n-3)$, respectively.

We now describe the simulation study.

- The prior density on β_2 (except for the benchmark) is such that

  $$\pi_{2,\omega}(\beta_2 \mid \sigma) = \frac{\lambda_2 n^{1/2}}{\sigma} g_2\left(\frac{\lambda_2 n^{1/2}}{\sigma}(\beta_2 - \mu_2)\right).$$

  We present the results for 4 choices of informative g_2: a standard normal distribution, a Student distribution, a LPTN distribution and a CTN distribution. We compare them with one another and to the non-informative prior.
- While keeping λ_2 fixed and equal to 1, we gradually increase μ_2 from 0 to 2. With this choice of λ_2, when $\mu_2 = 0$ the prior carries essentially the same information as the likelihood. We show the impact of more diffuse priors next. The results are presented in Figures 6 (a)–(b) and 7. Note that we observe similar results when considering a larger prior scaling, but we need to use an interval for μ_2 with a larger upper bound.
- While keeping μ_2 fixed and equal to 0.5 (to be able to appreciate a difference in location when the prior scaling conflicts), we gradually increase λ_2 from (nearly) 0 to 2. The results are presented in Figures 6 (c)–(d).

Figure 6 is used to compare the results produced by using different priors, while Figure 7 is used to show the impact of different choices of hyperparameters for the heavy-tailed priors, both in conflicting and non-conflicting situations. In Figure 6, we observe what has been explained before. Firstly, a Student prior resolves a conflict due to a prior location significantly different than that of the likelihood slower than a

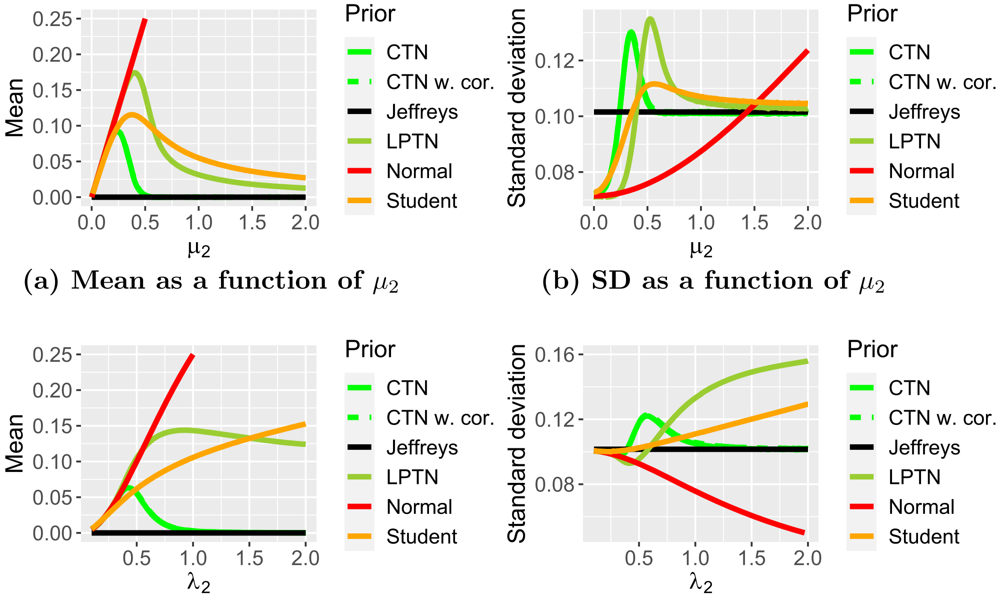
LPTN prior, i.e. the convergence towards the limiting posterior distribution is slower as $\mu_2 \to \infty$. Here, the limiting posterior resulting from a Student prior is not much different to that resulting from a LPTN prior; in both cases, the distribution of $\beta_2 \mid \sigma, \mathbf{y}$ is the same, but the distribution of $\sigma \mid \mathbf{y}$ is such that σ^2 follows an inverse-gamma and in the former case the shape and scale parameters are $(n - \gamma - 2)/2 = 47$ and $n/2$, respectively, whereas in the latter case, they are $(n - 2)/2 = 49$ and $n/2$, respectively. Similar arguments explain why, if we set g_2 to a CTN distribution and set the prior density of σ such that it is proportional to $\sigma\pi(\sigma)$ to correct for the trace asymptotically left by a CTN prior distribution, we obtain essentially the same estimates and standard deviations as if we did not correct for this trace and instead set the prior on σ to $\pi(\,\cdot\,)$ (the lines are on top of each other in Figure 6). In practice, one does the latter. Note that when we correct for that trace, we do it regardless of the values of μ_2 and λ_2, and therefore, for some values, we should not correct because the situations are non-conflicting. The correction is needed in the asymptotic regime, which is something theoretical, explaining why we did not discriminate.

In Figure 6, we also observe that using a Student or a LPTN prior is ineffective at resolving a conflict due to extremely small scalings (represented by $\lambda_2 \to \infty$), contrarily to using a CTN prior. A last point to note in Figure 6 is that, because the LPTN distribution is the most similar to the normal among the heavy-tailed distributions presented, using a LPTN prior translates into the closest results to those produced by using a normal one when there is no conflict, but also into the largest impact in the “gray” area, i.e. in between no conflict and clear conflict.

In Figure 7, we observe that increasing the level of similarity between an heavy-tailed prior and a normal prior (controlled through γ, ρ and ϱ for the Student, LPTN and CTN prior distributions, respectively) increases the threshold at which the Bayesian model starts to detect that the prior is conflicting (i.e. the point beyond which the impact starts to decrease) and thus increases the impact on the posterior distribution and estimate at this threshold. Our simulation study suggests that $\gamma = 4$, $\rho = 0.95$ and $\varrho = 0.98$ offer a good balance between great similarity with the standard normal (and thus great similarity in between the posterior distributions in the absence of conflict) and great capacity at detecting and resolving a conflict (due to a prior location significantly different than that of the likelihood for Student and LPTN priors). The impact of hyperparameters when there is a conflict due to small scalings is not shown because showing it is relevant only for CTN prior distributions, and the impact is similar as when the conflict is due to a prior location significantly different than that of the likelihood.

5 Conclusion

In this paper, we characterized the impact of using heavy-tailed alternatives to normal prior distributions for regression coefficients. This was achieved through a theoretical analysis under an asymptotic regime for which a conflicting situation becomes extreme and a simulation study, in Sections 3 and 4, respectively. The heavy-tailed alternatives are Student, LPTN and CTN prior distributions. With the results presented in hand,



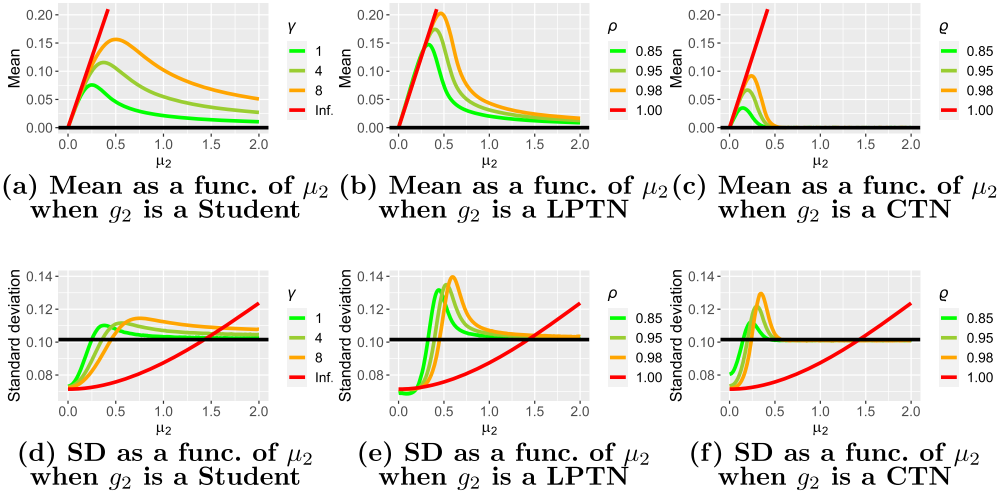
(a) Mean as a function of μ_2 **(b) SD as a function of μ_2**

(c) Mean as a function of λ_2 **(d) SD as a function of λ_2**

Figure 6: Impact on posterior means and standard deviations as μ_2 and λ_2 vary when a Jeffreys prior is used, i.e. $\pi_{2,\omega}(\beta_2 \mid \sigma) \propto 1$ (black line), and when g_2 is a standard normal (red line), a Student with $\gamma = 4$ degrees of freedom (orange line), a LPTN with $\rho = 0.95$ (dark green line), a CTN with $\varrho = 0.98$ and a CTN with $\varrho = 0.98$ but where the prior density of σ is proportional to $\sigma\pi(\sigma)$ (both lines are light green, one is dashed, while the other one not; they are on top of each other); here SD stands for standard deviation.

one is well equipped to decide which prior distributions to use for a Bayesian regression analysis. In summary, normal prior distributions can be used when one is confident that they will not be in conflict with the data to collect; otherwise, heavy-tailed alternatives should be employed. All heavy-tailed alternatives can be used in a situation of a potential conflict due to a prior location significantly different than that of the likelihood function. Using Student and CTN prior distributions has an impact on the posterior variability of all coefficients asymptotically as the conflict becomes extreme; the variability increases when using Student prior distributions, while it decreases when using CTN prior distributions. The impact is however small when the sample size is large relatively to the number of conflicting prior densities. Note that this is however only true for Student priors with small degrees of freedom. When the priors on the regression coefficients are such that one is exposed to potential conflicts due to prior scalings, the heavy-tailed alternative that is recommended is the CTN distribution.

The theoretical analysis performed in Section 3 was for linear regression. While there is no reason to believe that the results do not hold for other regression models, like GLMs, it would be interesting to prove similar results for such models to have a



(a) Mean as a func. of μ_2 when g_2 is a Student

(b) Mean as a func. of μ_2 when g_2 is a LPTN

(c) Mean as a func. of μ_2 when g_2 is a CTN

(d) SD as a func. of μ_2 when g_2 is a Student

(e) SD as a func. of μ_2 when g_2 is a LPTN

(f) SD as a func. of μ_2 when g_2 is a CTN

Figure 7: Impact on posterior means and standard deviations as μ_2 varies when: (a) and (d) g_2 is a Student, for different values of γ (Inf. represents the standard normal); (b) and (e) g_2 is a LPTN, for different values of ρ (1.00 represents the standard normal); (c) and (f) g_2 is a CTN, for different values of ϱ (1.00 represents the standard normal); here SD stands for standard deviation and the black lines are again the results for the Jeffreys prior.

confirmation and to have access to precise statements describing the conditions under which the results hold.

Supplementary Material

Robustness against conflicting prior information in regression – supplementary material (DOI: 10.1214/22-BA1330SUPP; .pdf).

References

Bunke, O. and Milhaud, X. (1998). “Asymptotic behavior of Bayes estimates under possibly incorrect models.” *The Annals of Statistics*, 26(2): 617–644. MR1626075. doi: `https://doi.org/10.1214/aos/1028144851`. 851

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). “The horseshoe estimator for sparse signals.” *Biometrika*, 97(2): 465–480. MR2650751. doi: `https://doi.org/10.1093/biomet/asq017`. 842

Dawid, A. P. (1973). “Posterior expectations for large observations.” *Biometrika*, 60(3): 664–667. MR0336889. doi: `https://doi.org/10.1093/biomet/60.3.664`. 844

de Finetti, B. (1961). "The Bayesian approach to the rejection of outliers." In *Proceedings of the fourth Berkeley Symposium on Probability and Statistics*, volume 1, 199–210. University of California Press Berkeley. MR0133935. 844

Desgagné, A. (2015). "Robustness to Outliers in Location-Scale Parameter Model using Log-Regularly Varying Distributions." *The Annals of Statistics*, 43(4): 1568–1595. MR3357871. doi: https://doi.org/10.1214/15-AOS1316. 841, 844, 845, 848

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. 844

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). "Hybrid Monte Carlo." *Physics Letters B*, 195(2): 216–222. MR3960671. doi: https://doi.org/10.1016/0370-2693(87)91197-x. 848

Gagnon, P. (2022). "Robustness against conflicting prior information in regression – supplementary material." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1330SUPP. 845

Gagnon, P., Bédard, M., and Desgagné, A. (2021). "An automatic robust Bayesian approach to principal component regression." *Journal of Applied Statistics*, 48(1): 84–104. arXiv:1711.06341. MR4183269. doi: https://doi.org/10.1080/02664763.2019.1710478. 844, 847

Gagnon, P., Desgagné, A., and Bédard, M. (2020). "A New Bayesian Approach to Robustness Against Outliers in Linear Regression." *Bayesian Analysis*, 15(2): 389–414. MR4078719. doi: https://doi.org/10.1214/19-BA1157. 844, 847, 854, 855

Gagnon, P. and Hayashi, Y. (2022). "Theoretical properties of Bayesian Student-t linear regression." arXiv:2204.02299. 844

Hamura, Y., Irie, K., and Sugasawa, S. (2022). "Log-regularly varying scale mixture of normals for robust regression." *Computational Statistics & Data Analysis*, 173: 107517. MR4418923. doi: https://doi.org/10.1016/j.csda.2022.107517. 844

Hill, B. M. (1974). "On coherence, inadmissibility and inference about many parameters in the theory of least squares." In *Studies in Bayesian econometrics and statistics: In honor of Leonard J. Savage*, 555–584. Amsterdam: North-Holland. MR0431488. 844

Johnstone, I. M. and Silverman, B. W. (2004). "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences." *The Annals of Statistics*, 32(4): 1594–1649. MR2089135. doi: https://doi.org/10.1214/009053604000000030. 842

Lindley, D. V. (1968). "The choice of variables in multiple regression." *Journal of the Royal Statistical Society: Series B (Methodology)*, 30(1): 31–53. MR0231492. 844

Mutlu, K., Çankaya, E., and Arslan, O. (2019). "Robust Bayesian Regression Analysis Using Ramsay-Novick Distributed Errors with Student-t Prior." *Communications Faculty of Sciences University of Ankara Series A1: Mathematics and Statistics*, 68(1): 602–618. MR3827540. doi: https://doi.org/10.31801/cfsuasmas.441096. 846

O'Hagan, A. (1979). "On outlier rejection phenomena in Bayes inference." *Journal of the Royal Statistical Society: Series B (Methodology)*, 41(3): 358–367. MR0557598. 844

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. 841, 844

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 842

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association*, 92(437): 179–191. MR1436107. doi: https://doi.org/10.2307/2291462. 842, 857

Roberts, G. O. and Tweedie, R. L. (1996). "Exponential convergence of Langevin distributions and their discrete approximations." *Bernoulli*, 2(4): 341–363. MR1440273. doi: https://doi.org/10.2307/3318418. 848

Scheffé, H. (1947). "A Useful Convergence Theorem for Probability Distributions." *The Annals of Mathematical Statistics*, 434–438. MR0021585. doi: https://doi.org/10.1214/aoms/1177730390. 854

West, M. (1984). "Outlier Models and Prior Distributions in Bayesian Linear Regression." *Journal of the Royal Statistical Society: Series B (Methodology)*, 46(3): 431–439. MR0790630. 842, 846, 857

Acknowledgments

The author thanks two anonymous referees and an associate editor for helpful suggestions that led to an improved manuscript.