Brazilian Journal of Probability and Statistics
2019, Vol. 33, No. 2, 205–221
https://doi.org/10.1214/17-BJPS385

Bayesian robustness to outliers in linear regression and ratio estimation

Alain Desgagné[a] **and Philippe Gagnon**[b]

[a]*Université du Québec à Montréal*
[b]*Université de Montréal*

Abstract. Whole robustness is a nice property to have for statistical models. It implies that the impact of outliers gradually vanishes as they approach plus or minus infinity. So far, the Bayesian literature provides results that ensure whole robustness for the location-scale model. In this paper, we make two contributions. First, we generalise the results to attain whole robustness in simple linear regression through the origin, which is a necessary step towards results for general linear regression models. We allow the variance of the error term to depend on the explanatory variable. This flexibility leads to the second contribution: we provide a simple Bayesian approach to robustly estimate finite population means and ratios. The strategy to attain whole robustness is simple since it lies in replacing the traditional normal assumption on the error term by a super heavy-tailed distribution assumption. As a result, users can estimate the parameters as usual, using the posterior distribution.

1 Introduction

Conflicting sources of information may contaminate the inference arising from statistical analysis. The conflicting information may come from outliers and also prior misidentification. In this paper, we focus on robustness with respect to outliers in a Bayesian simple linear regression model through the origin. We say that a conflict occurs when a group of observations produces a rather different inference than that proposed by the bulk of the data and the prior. Light-tailed distribution assumptions on the error term can lead to an undesirable compromise where the posterior distribution concentrates on an area that is not supported by any source of information. We believe that the appropriate way to address the problem is to limit the influence of outliers in order to obtain conclusions consistent with the majority of the observations.

Box and Tiao (1968) were the first to introduce a robust Bayesian linear regression model. They proposed to assume that the distribution of the error term is a mixture of two normals with 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

Key words and phrases. Built-in robustness, simple linear regression, ratio estimator, finite populations, population means, super heavy-tailed distributions.
Received December 2016; accepted October 2017.

scale mixtures of normals, which include the Student distribution. More recently, Peña, Zamar and Yan (2009) introduced a different robust Bayesian method where each observation has a weight decreasing with the distance between this observation and most of the data. They proved that the Kullback–Leibler divergence from the posterior arising from the nonoutliers only to the posterior arising from the sample containing outliers is bounded.

So far, the literature only provides solutions to attain whole robustness for the estimation of the slope in the model of regression through the origin (e.g. if we assume that the error term has a Student distribution instead of a normal, see the results of Andrade and O'Hagan (2011) in a context of location-scale model). However, only partial robustness is reached for the estimation of the scale parameter of the error term. Partial robustness means that the outliers have a significant but limited influence on the inference, as the conflict grows infinitely. In this paper, we go a step further: we attain whole robustness to outliers for both the slope and scale parameters, in the sense that the impact of outliers gradually vanishes as they approach plus or minus infinity. To achieve this, we generalise the results of Desgagné (2015), which ensure whole robustness for both parameters of the location-scale model simultaneously, to the simple linear regression model through the origin. Our work is thus aligned with the *theory of conflict resolution in Bayesian statistics*, as described by O'Hagan and Pericchi (2012) in their extensive literature review on that topic.

The strategy to attain whole robustness for all parameters is, instead of assuming the traditional normality of the errors in the model, to assume that they have a super heavy-tailed distribution. The general model (with no specific distribution assumption on the error term) is described in Section 2.1. The class of super heavy-tailed distributions that we consider, which are log-regularly varying distributions, is presented in Section 2.2. When assuming a super heavy-tailed distribution on the error term, the resulting model is characterised by its built-in robustness that resolves conflicts in a sensitive and automatic way, as stated in our robustness results given in Section 2.3. The main result is the convergence of the posterior distribution towards the posterior arising from the nonoutliers only, when the outliers approach plus or minus infinity. Although our results are Bayesian analysis-oriented, they reach beyond this paradigm through the robustness of the likelihood function, and therefore, of both slope and scale maximum likelihood parameter estimation. These are the results that ensure that whole robustness is reached for the considered model.

We believe our work will eventually lead to whole robustness results for the estimation of the parameters of the usual multiple linear regression model, which will in turn allow to introduce Bayesian robust ANOVA and t-test procedures. In fact, a preliminary numerical investigation suggests that similar results to those presented in this paper hold for multiple linear regressions. However, precise conditions and results will need to be specified. This can be achieved by the (non-trivial) extension

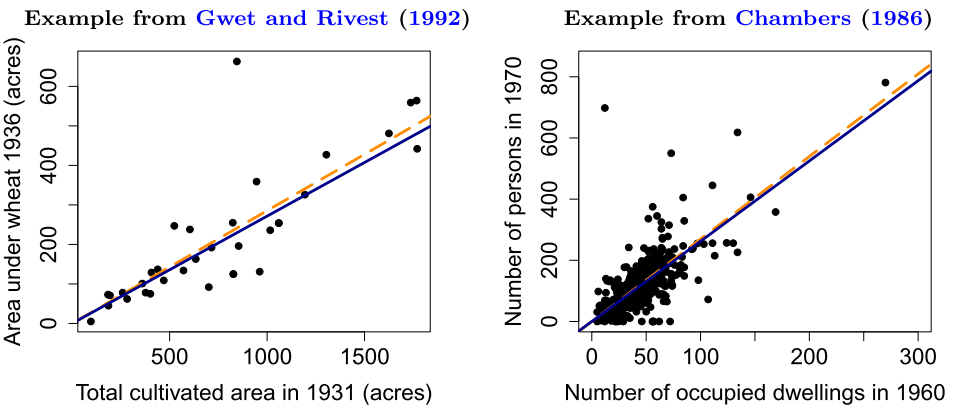

Figure 1 *Example of data sets containing outliers with slope estimates under normal* (*orange dashed line*), *and super heavy-tailed* (*blue solid line*) *distribution assumptions*; *the data sets are provided in the supplementary material* (*Desgagné and Gagnon* (2019)).

of the proof presented in the following for the simple linear regression through the origin.

In addition to representing a crucial step towards whole robustness for the more general case of multiple linear regressions, whole robustness for the simple linear regression through the origin finds an important application in the estimation of ratios and finite population means. As shown in Figure 1, one may encounter the presence of outliers in achieving this task. In Gwet and Rivest (1992), the ratio aimed to be estimated was the area under wheat in 1936 to the total cultivated area in 1931 in a given administrative geographical unit of Uttar Pradesh state in India. In Chambers (1986), it was the total population in 1970 in East Baltimore to the number of occupied dwelling in 1960 in the same area. In Section 3, we illustrate the relevance of our robust approach through analyses in economic contexts. More precisely, the following contexts are considered: robust estimation of the personal disposable income per capita and of the average weekly household expenditure on food (using the ratio estimator). In Section 3, we also detail the link between simple linear regression through the origin and finite population sampling, and present a simulation study. In all analyses, our approach is compared with the nonrobust (with the normal assumption) and partially robust (with the Student distribution assumption) approaches. It is showed that our model performs as well as the nonrobust and the partially robust models in absence of outliers, in addition to being completely robust. It indicates that, by only changing the assumption on the error term, we obtain adequate estimates in absence or presence of outliers. These estimates are computed as usual from the posterior distribution.

2 Resolution of conflicts in simple linear regression through the origin

2.1 Model

(i) Let $Y_1, \ldots, Y_n \in \mathbb{R}$ be n random variables and $x_1, \ldots, x_n \in \mathbb{R}\backslash\{0\}$ be n known constants, where $n > 2$ is assumed to be known. We assume that

$$Y_i = \beta x_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n \in \mathbb{R}$ and $\beta \in \mathbb{R}$ are $n + 1$ conditionally independent random variables given $\sigma > 0$ with a conditional density for ε_i given by

$$\varepsilon_i \mid \beta, \sigma \overset{\mathcal{D}}{=} \varepsilon_i \mid \sigma \overset{\mathcal{D}}{\sim} \frac{1}{\sigma |x_i|^\theta} f\left(\frac{\varepsilon_i}{\sigma |x_i|^\theta}\right), \qquad i = 1, \ldots, n,$$

$\theta \in \mathbb{R}$ being a known constant.

(ii) We assume that f is a strictly positive continuous probability density function on $\mathbb{R}$ that is symmetric with respect to the origin, and that is such that both tails of $|z| f(z)$ are monotonic, which implies that the tails of $f(z)$ are also monotonic. The density f can have parameters, for example, a shape parameter; however, their value is assumed to be known.

(iii) We assume that the prior of β and σ, denoted $\pi(\beta, \sigma)$, is bounded on $\sigma > 1$, and is such that $\pi(\beta, \sigma)/(1/\sigma)$ is bounded on $0 < \sigma \leq 1$, for all $\beta \in \mathbb{R}$. Together, these assumptions are equivalent to: $\pi(\beta, \sigma)/\max(1, 1/\sigma)$ is bounded on $\sigma > 0$. A large variety of priors fit within this assumed structure; for instance, this is the case for all proper densities. In addition, non-informative priors such as $\pi(\beta, \sigma) \propto 1/\sigma$, the usual one for this type of random variables, and $\pi(\beta, \sigma) \propto 1$ satisfy these assumptions.

From this perspective, $x_1, \ldots, x_n$ represent observations of the explanatory variable, the dependent variable and the error term are respectively, represented by the continuous random variables $Y_1, \ldots, Y_n$ and $\varepsilon_1, \ldots, \varepsilon_n$, and the parameter β represents the slope of the regression line. Note that no assumptions are made on the explanatory variable, except that the value 0 cannot be observed.

The scale of the distribution of the error term is $\sigma |x_i|^\theta$ and, therefore, the variability of the errors increases (decreases) as x_i moves away from 0 when $\theta > 0$ ($\theta < 0$). This model can thus be used in a context of heteroscedasticity. When the classical framework is considered, that is, a frequentist setting with the assumption that f is the standard normal density, $\sigma |x_i|^\theta$ also represents the standard deviation of the error ε_i. In this situation, the maximum likelihood estimator of β is the weighted average of the y_i/x_i given by $\hat{\beta} = \sum_{i=1}^n w_i (y_i/x_i)$, where $w_i = |x_i|^{2(1-\theta)} / \sum_{j=1}^n |x_j|^{2(1-\theta)}$.

An important drawback of the classical framework is that outliers have a significant impact on the estimation, due to the normal assumption. In this paper, we study robustness of the estimation of β and σ. The objective is to find sufficient conditions to attain whole robustness. The nature of the results presented in

Section 2.3 is asymptotic, in the sense that some y_i's approach $+\infty$ or $-\infty$. The known vector $\mathbf{x_n} := (x_1, \ldots, x_n)$ is considered as fixed. In Section 3.1, we explain that studying this theoretical framework is sufficient to attain, in practice, robustness against any type of outliers (i.e. outliers because of their extreme x value, extreme y value, or both).

Among the n observations of $Y_1, \ldots, Y_n$, denoted by $\mathbf{y_n}$, we assume that $k > 2$ of them, denoted by the vector $\mathbf{y_k}$, form a group of nonoutlying observations, m of them are considered as "negative slope outliers", with relatively small (large) values of y_i when x_i is positive (negative), and p of them are considered as "positive slope outliers", with relatively large (small) values of y_i when x_i is positive (negative), with $k + m + p = n$. Note that we use the letter m for "minus" because the related outliers attract the slope towards negative values, and analogously, we use the letter p for "positive". For $i = 1, \ldots, n$, we define the binary functions k_i, m_i and p_i as follows: if y_i is a nonoutlying value, $k_i = 1$; if it is a negative slope outlier, $m_i = 1$ and if it is a positive slope outlier, $p_i = 1$. These functions take the value of 0 otherwise. Therefore, we have $k_i + m_i + p_i = 1$ for $i = 1, \ldots, n$, with $\sum_{i=1}^n k_i = k$, $\sum_{i=1}^n m_i = m$ and $\sum_{i=1}^n p_i = p$. We assume that each outlier approaches $-\infty$ or $+\infty$ at its own specific rate, to the extent that the ratio of two outliers is bounded. More precisely, we assume that $y_i = a_i + b_i\omega$, for $i = 1, \ldots, n$, where a_i and b_i are constants such that $a_i \in \mathbb{R}$ and

(i) $b_i = 0$ if $k_i = 1$,
(ii) $b_i < 0$ if y_i is "small", that is if $x_i < 0$, $p_i = 1$ or $x_i > 0$, $m_i = 1$,
(iii) $b_i > 0$ if y_i is "large", that is if $x_i < 0$, $m_i = 1$ or $x_i > 0$, $p_i = 1$,

and we let $\omega \to \infty$.

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

$$\pi(\beta, \sigma \mid \mathbf{y_n}) = \big[m(\mathbf{y_n})\big]^{-1}\pi(\beta, \sigma)\prod_{i=1}^{n}\frac{1}{\sigma|x_i|^\theta}f\left(\frac{y_i - \beta x_i}{\sigma|x_i|^\theta}\right), \qquad \beta \in \mathbb{R}, \sigma > 0.$$

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

$$\pi(\beta, \sigma \mid \mathbf{y_k}) = \big[m(\mathbf{y_k})\big]^{-1}\pi(\beta, \sigma)\prod_{i=1}^{n}\left[\frac{1}{\sigma|x_i|^\theta}f\left(\frac{y_i - \beta x_i}{\sigma|x_i|^\theta}\right)\right]^{k_i}, \qquad \beta \in \mathbb{R}, \sigma > 0.$$

Note that if the prior $\pi(\beta, \sigma)$ is proportional to 1, the likelihood functions, given by the product term in the posteriors above, can also be expressed as follows:

$$\mathcal{L}(\beta, \sigma \mid \mathbf{y_n}) = m(\mathbf{y_n})\pi(\beta, \sigma \mid \mathbf{y_n}) \quad \text{and} \quad \mathcal{L}(\beta, \sigma \mid \mathbf{y_k}) = m(\mathbf{y_k})\pi(\beta, \sigma \mid \mathbf{y_k}). \quad (1)$$

Proposition 1. *Considering the Bayesian context given in Section 2.1, the joint posterior densities $\pi(\beta, \sigma \mid \mathbf{y_k})$ and $\pi(\beta, \sigma \mid \mathbf{y_n})$ are proper.*

The proof of Proposition 1 can be found in the supplementary material.

2.2 Log-regularly varying distributions

As mentioned in the Introduction, our approach to attain robustness is to replace the traditional normal assumption on the error term by a log-regularly varying distribution assumption. The definition of such a distribution is now presented.

Definition 1 (Log-regularly varying distribution). A random variable Z with a symmetric density $f(z)$ is said to have a log-regularly varying distribution with index $\rho \geq 1$ if $zf(z) \in L_\rho(\infty)$, meaning that $zf(z)$ is *log-regularly varying* at ∞ with index $\rho \geq 1$.

Log-regularly varying functions is an interesting class of functions with useful properties for robustness. By definition, they are such that $g \in L_\rho(\infty)$ if $g(z^\nu)/g(z)$ converges towards $\nu^{-\rho}$ uniformly in any set $\nu \in [1/\tau, \tau]$ (for any $\tau \geq 1$) as $z \to \infty$, where $\rho \in \mathbb{R}$. This implies that for any $\rho \in \mathbb{R}$, we have $g \in L_\rho(\infty)$ if and only if there exists a constant $A > 1$ and a function $s \in L_0(\infty)$ (which is called a log-slowly varying function) such that for $z \geq A$, g can be written as $g(z) = (\log z)^{-\rho} s(z)$. An example of log-regularly varying distributions is presented in Section 3.1. The purpose of this section was to provide an overview of the tail behaviour of such distributions. For more information on log-regularly varying distributions, we refer the reader to Desgagné (2013) and Desgagné (2015).

2.3 Resolution of conflicts

The results of robustness are now given in Theorem 1.

Theorem 1. *Consider the model and the context described in Section 2.1. If we assume that*

(i) $zf(z) \in L_\rho(\infty)$, *with* $\rho \geq 1$ (*i.e. that* f *is a log-regularly varying distribution*),
(ii) $k > \max(m, p)$ (*i.e. that both the negative and positive slope outliers are fewer than the nonoutliers*),

then, recalling that $y_i = a_i + b_i\omega$ *with* $b_i = 0$ *for the nonoutliers and* $b_i \neq 0$ *for the outliers, we obtain the following results*:

(a)

$$\lim_{\omega \to \infty} \frac{m(\mathbf{y_n})}{\prod_{i=1}^{n} [f(y_i)]^{m_i + p_i}} = m(\mathbf{y_k}),$$

(b)

$$\lim_{\omega \to \infty} \pi(\beta, \sigma \mid \mathbf{y_n}) = \pi(\beta, \sigma \mid \mathbf{y_k}),$$

uniformly on $(\beta, \sigma) \in [-\lambda, \lambda] \times [1/\tau, \tau]$, *for any* $\lambda \geq 0$ *and* $\tau \geq 1$,

(c)

$$\lim_{\omega\to\infty}\int_0^\infty\int_{-\infty}^\infty \left|\pi(\beta,\sigma\mid \mathbf{y_n})-\pi(\beta,\sigma\mid \mathbf{y_k})\right| d\beta\, d\sigma = 0,$$

(d) *as* $\omega\to\infty$,

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

and in particular

$$\beta\mid \mathbf{y_n}\xrightarrow{\mathcal{D}}\beta\mid \mathbf{y_k}\quad \textit{and}\quad \sigma\mid \mathbf{y_n}\xrightarrow{\mathcal{D}}\sigma\mid \mathbf{y_k},$$

(e)

$$\lim_{\omega\to\infty}\big[m(\mathbf{y_k})/m(\mathbf{y_n})\big]\mathcal{L}(\beta,\sigma\mid \mathbf{y_n})=\mathcal{L}(\beta,\sigma\mid \mathbf{y_k}),$$

uniformly on $(\beta,\sigma)\in[-\lambda,\lambda]\times[1/\tau,\tau]$, *for any* $\lambda\geq 0$ *and* $\tau\geq 1$.

The proof of Theorem 1 can be found in the supplementary material. Note that, when $x_1=\cdots=x_n=1$, the simple linear regression model through the origin becomes the location-scale model, and this highlights the fact that our results generalise those of Desgagné (2015).

Theorem 1 is particularly appealing for its simplicity, and therefore, for its practical use. Indeed, condition (i) only indicates that modelling must be done using a density f with sufficiently heavy tails, specifically with a log-regularly varying distribution (see Definition 1). For that purpose, Desgagné (2015) introduced the family of log-Pareto-tailed symmetric distributions, which belongs to the family of log-regularly varying distributions and therefore satisfies condition (i). This new family includes, for instance, piecewise densities constructed from well-known symmetric densities like the normal, uniform or Student by replacing their extremities by log-Pareto tails, that is, tails that behave like $(1/|z|)(\log|z|)^{-\phi}$ with $\phi>1$. A special case of log-Pareto-tailed symmetric distributions, called the log-Pareto-tailed standard normal (LPTN) distribution with parameters $\alpha>1$ and $\phi>1$, is given in Section 3.1. It exactly matches the standard normal on the interval $[-\alpha,\alpha]$, with log-Pareto tails. This is the super heavy-tailed distribution that we use in our numerical analyses. Note that we can also construct symmetric densities with log-Pareto tails that are not piecewise through transformations of the Pareto distribution. For instance, from a Pareto random variable Y with density $g(y)=\phi\theta^\phi y^{-(\phi+1)}$, $y>\theta$, we can make the change of variable $|Z|=e^Y-e^\theta \Leftrightarrow Y=\log(|Z|+e^\theta)$ to obtain a double-log-Pareto distribution with density

$$f(z)=(1/2)\phi\theta^\phi\big(|z|+e^\theta\big)^{-1}\big[\log\big(|z|+e^\theta\big)\big]^{-(\phi+1)},$$
$$-\infty<z<\infty,\theta>0,\phi>0.$$

Condition (ii) indicates that both the negative and positive slope outliers must be fewer than the nonoutlying observations, i.e. $m < k$ and $p < k$. In other words, the nonoutlying observations must form the largest group. For instance, with a sample of size $n = 25$, the model rejects up to 16 outliers if they are split in $m = 8$ negative and $p = 8$ positive slope outliers, which leaves $k = 9$ nonoutliers. At the other end of the spectrum, in the situation where all outliers are of the same type, for instance all positive slope outliers (which implies that $m = 0$), the model rejects up to $p = 12$ outliers, which leaves $k = 13$ nonoutliers. Numerical simulations seem to confirm our expectation that a larger difference between k and $\max(m, p)$ results in a more rapid rejection of the outliers.

The breakdown point is generally defined as the largest proportion of outliers that an estimator can handle. In our situation, for a sample size of n, the condition $k > \max(m, p)$ translates into a breakdown point of $\lfloor (n-1)/2 \rfloor / n$, that is the integer part of $(n-1)/2$ divided by n, if we consider only positive slope outliers (or only negative slope outliers). As $n \to \infty$, the breakdown point converges to 0.5, usually considered as the maximum desired value.

Not only do the conditions of Theorem 1 are simple and intuitive, the results are also easy to interpret. The asymptotic behaviour of the marginal $m(\mathbf{y_n})$ is described by result (a). While this result is more of theoretical interest, it is the cornerstone of this robustness theory; it leads to results (b) to (e), which are more practical. Result (b) indicates that the posterior density, arising from the whole sample, converges towards the posterior density arising from the nonoutliers only, uniformly in any set $(\beta, \sigma) \in [-\lambda, \lambda] \times [1/\tau, \tau]$. The impact of the outliers then gradually decreases to nothing as they approach plus or minus infinity.

Result (b) leads to result (c): the convergence in L_1 of the posterior density, arising from the whole sample, towards the posterior density arising from the nonoutlying observations only. This last result implies the following convergence: $\mathbb{P}(\beta, \sigma \in E \mid \mathbf{y_n}) \to \mathbb{P}(\beta, \sigma \in E \mid \mathbf{y_k})$ as $\omega \to \infty$, uniformly for all rectangles $E \in \mathbb{R} \times \mathbb{R}^+$. This result is slightly stronger than convergence in distribution (result (d)) which requires only pointwise convergence. Then, the convergence of the posterior marginal distributions is directly obtained. Therefore, any estimation of β and σ based on posterior quantiles (for example posterior medians and Bayesian credible intervals) is robust to outliers. Note that results (a) to (d) are also valid if we assume that $n \geq 2, k \geq 2$ (instead of $n > 2, k > 2$), provided that we assume that $\sigma \pi(\beta, \sigma)$ is bounded (instead of $\min(\sigma, 1)\pi(\beta, \sigma)$ is bounded).

Result (e) indicates that, for a given sample, the likelihood (up to a multiplicative constant that does not depend on β and σ) converges to the likelihood arising from the nonoutliers only, uniformly in any set $(\beta, \sigma) \in E$, where $E = [-\lambda, \lambda] \times [1/\tau, \tau]$. Consequently, the maximum of $\mathcal{L}(\beta, \sigma \mid \mathbf{y_n})$ thus converges to the maximum of $\mathcal{L}(\beta, \sigma \mid \mathbf{y_k})$ on the set E and, therefore the maximum likelihood estimate also converges, as $\omega \to \infty$. Note that, using results (b) to (d), we know that, for both $\pi(\beta, \sigma \mid \mathbf{y_k})$ and $\pi(\beta, \sigma \mid \mathbf{y_n})$, the volume on E^c over the volume on E converges to 0 as λ and τ increase; this relation holds in particular if $\pi(\beta, \sigma) \propto 1$ and, in this case, the posterior is proportional to the likelihood.

3 Finite population means and ratios

To use the model described in Section 2.1, users have to set the value of θ. Different particular values lead to interesting special cases. For instance, when $\theta = 0$, the resulting model is the classical homoscedastic model, with $\text{Var}(\varepsilon_i) = \sigma^2$ and $\hat{\beta} = \sum_{i=1}^{n} x_i y_i / \sum_{j=1}^{n} x_j^2$, considering the classical framework. When $\theta = 1$, the estimator of β is the unweighted mean of the y_i/x_i, that is $\hat{\beta} = (1/n) \sum_{i=1}^{n} y_i/x_i$. Probably the most interesting special case results from $\theta = 1/2$ and $x_i > 0$ for all i. Indeed, considering again the classical framework, the estimator of β is $\hat{\beta} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$, which is commonly used to estimate the following finite population ratio: $\sum_{i=1}^{N} y_i / \sum_{i=1}^{N} x_i$, where y_i and x_i are measures of the variable of interest and of the auxiliary variable on unit i, respectively, and N is the population size. The estimator $\hat{\beta} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$ is also used to estimate the finite population mean μ_y of a variable of interest y using auxiliary information of a variable x as follows: $\hat{\mu}_y = \hat{\beta} \times \mu_x$, where μ_x is the known population mean of x. This last estimator is known as the ratio estimator and to be more accurate than the simple location model when the variable of interest is correlated with the auxiliary variable. Therefore, robust estimators of β lead to robust estimators of finite population means and ratios. To our knowledge, Gwet and Rivest (1992) introduced the first frequentist outlier resistant alternatives to the ratio estimator, using well known *M*- (Huber (1973)) and *GM*- (Mallows (1975)) estimators. Their research was inspired by the work of Chambers (1986), the first author to use regression *M*-estimators in survey sampling.

In Section 3.1, we present real-life situations in which ratio estimation is useful, while illustrating the theoretical results of Theorem 1. First, in a context of estimation of personal disposable income (PDI) per capita, we show that, when we artificially move an observation, its impact on the estimation grows until it reaches a certain threshold. Beyond this threshold, the impact decreases to nothing as the observation approaches plus or minus infinity. Second, a more traditional Bayesian analysis is made, in which we study the proportion of income spent on food. More precisely, we present the posterior distributions, with particular emphasis on the impact of outliers, and we compute various estimates from the posteriors. In Section 3.2, again in a context of finite population sampling, a simulation study is conducted to evaluate the accuracy of the estimates arising from our model. In all analyses, we compare its performance with those of the nonrobust (the model with the normal assumption) and partially robust (the model with the Student distribution assumption) models. As mentioned in Section 1, the model of Box and Tiao (1968) can be viewed as a special case of the partially robust model. We therefore omit the comparison with their model. In the simulation study, we also consider the following frequentist competitors: the *M*- and *S*- (Rousseeuw and Yohai (1984)) estimators. R functions that are used for the computations are provided in Section 2 of the supplementary material.

3.1 Illustration of the results of Theorem 1

In the first context, we are interested in the estimation of the PDI per capita when the available data are the total disposable income (y_i) for n households (in this analysis $n = 20$), and the number of individuals (x_i) in each of these households. The data are presented in Table 1. The PDI per capita, which is a population mean per individual, would be directly computed by $\sum_{i=1}^{N} y_i / \sum_{i=1}^{N} x_i$ (where N is the number of households in the population) if the information was available for all the households. We therefore use the simple linear regression model through the origin with $\theta = 1/2$ to estimate this ratio (see Section 2.1 for details about the model).

In order to illustrate the threshold feature, an observation is randomly chosen (in this analysis, it is the 11th observation), and y_{11} is gradually moved from the value 100 (a nonoutlier) to 385 (a large outlier), while $x_{11} = 3$ remains fixed. The parameters β and σ are estimated for each data set related to a different value of y_{11} using maximum *a posteriori* probability (MAP) estimation with a prior proportional to 1 (which corresponds to maximum likelihood estimation). This process is performed under three models, each corresponding to a different assumption on f: a standard normal density (in this case, $\hat{\beta} = \sum_{i=1}^{20} y_i / \sum_{i=1}^{20} x_i$, the classical ratio estimator), a Student density (the partially robust model) or a LPTN density (our robust model). The results are presented in Figure 2.

The inference is clearly not robust when it is assumed that the error has a normal distribution (orange dashed line) since the values of the point estimates of β and σ increase with y_{11}. Regarding the second model, the degrees of freedom of the heavy-tailed Student distribution have been arbitrarily set to 10 and a known scale parameter of 0.88 has been added to this distribution in order to have the same 2.5th and 97.5th percentiles as the standard normal. The estimation of β is robust as the impact of the outlier slowly decreases after a certain threshold. However, the estimation of σ is only partially robust, that is, the impact of the outlier is limited, but does not decrease when the outlying value increases. For the last model, we set α of the LPTN to 1.96 so that this distribution matches the standard normal on the interval $[-1.96, 1.96]$, implying that both distributions have the same 2.5th and 97.5th percentiles. Therefore, all three distributions studied in this section have 95% of their mass in the interval $[-1.96, 1.96]$. The other parameter of the LPTN ϕ is equal to 4.08 according to the procedure described in Section 4 of

Table 1 *Total disposable income for household i in thousands of dollars* (y_i) *and the number of individuals in household i* (x_i), *for* $i = 1, \ldots, 20$

y_i	20.8	9.6	38.6	74.1	108.8	98.7	44.8	77.2	93.2	107.2
x_i	1.0	1.0	2.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
y_i	y_{11}	93.6	113.7	123.5	93.5	148.1	147.1	154.0	149.5	173.5
x_i	3.0	4.0	4.0	4.0	4.0	5.0	5.0	5.0	6.0	6.0

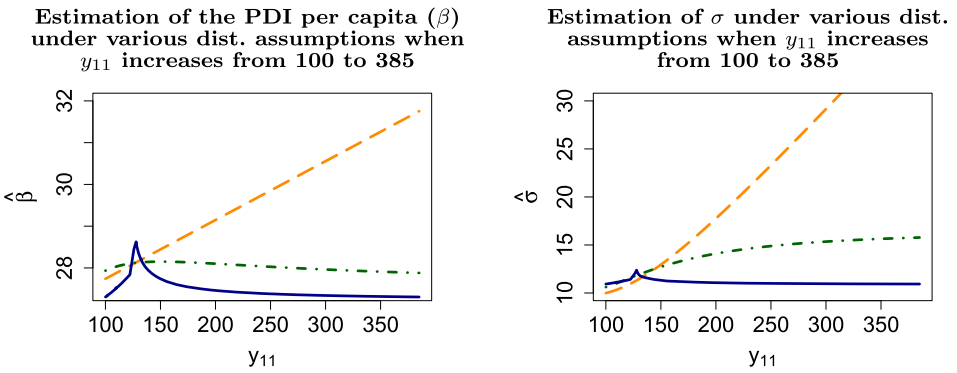

Figure 2 *Estimation of the PDI per capita* (β) *and* σ *when* y_{11} *increases from* 100 *to* 385 *under three different assumptions on* f: *standard normal density* (*orange dashed line*), *Student density* (*green dot-dashed line*) *and LPTN density* (*blue solid line*).

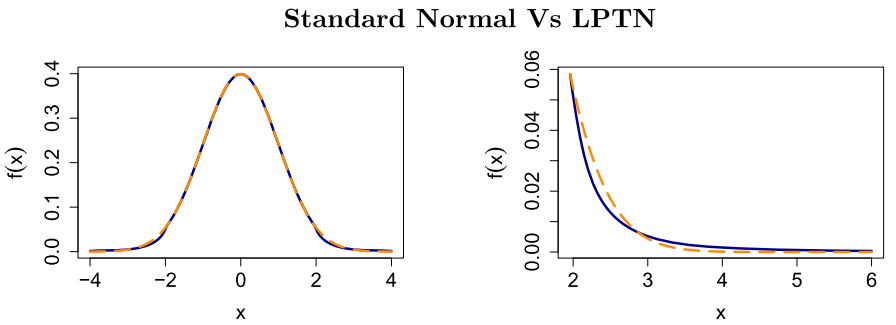

Figure 3 *Densities of the standard normal* (*orange dashed line*) *and of the LPTN with* $\alpha = 1.96$ *and* $\phi = 4.08$ (*blue solid line*).

Desgagné (2015) (this procedure ensures that f is continuous and a probability density function). The density of the LPTN distribution, depicted in Figure 3, is given by

$$f(x) = \begin{cases} \varphi(x) & \text{if } |x| \leq \alpha \text{ (the standard normal part),} \\ \varphi(\alpha)\big(\alpha/|x|\big)\big(\log\alpha/\log|x|\big)^{\phi} & \text{if } |x| > \alpha \text{ (the log-Pareto tails),} \end{cases} \tag{2}$$

where $\phi = 1 + 2\varphi(\alpha)\alpha\log(\alpha)/(1-q)$ and $q = \Phi(\alpha) - \Phi(-\alpha)$, φ and Φ being the probability density function and cumulative distribution function of the standard normal distribution, respectively.

For our robust model, it can be seen that y_{11} has an increasing impact on the estimation until this observation reaches a threshold. In this analysis, the threshold is around $y_{11} = 127.9$, and based on the data set with $y_{11} = 127.9$, $\hat{\beta} = 28.6$ and $\hat{\sigma} = 12.4$, which is interpreted as: the personal disposable income per capita is approximately 28,600. Beyond this threshold, the impact of the outlier gradually

decreases to nothing as the conflict grows infinitely. The point estimates converge towards 27.1 for β and 10.8 for σ, which are the point estimates when (x_{11}, y_{11}) is excluded from the sample. Whole robustness is therefore attained for both β and σ. Note that an increase in the value of the parameter α would result in an increase in the value of the threshold. Setting $\alpha = 1.96$ seems to be suitable for practical use.

In the second context, we are interested in the estimation of the proportion of weekly income spent on food for a population, when data are available per household. If the information was available for the population, we would directly compute the proportion by $\sum_{i=1}^{N} y_i / \sum_{i=1}^{N} x_i$, where N is the number of households in the population, and y_i and x_i are respectively the weekly expenditure on food and the weekly income, for household i. This ratio can thus be approximated using the simple linear regression through the origin with $\theta = 1/2$, and again, we compare our robust model with the nonrobust and partially robust models. We use the same Student and LPTN distributions as in the first context above, but we set the prior $\pi(\beta, \sigma) \propto 1/\sigma$. A Markov chain Monte Carlo (MCMC) method is implemented for the estimation (see Section 2 of the supplementary material for the R functions). It is run for 10,000,000 iterations.

Note that the ratio $\sum_{i=1}^{N} y_i / \sum_{i=1}^{N} x_i$ can be viewed as the following weighted average: $\sum_{i=1}^{N} w_i (y_i / x_i)$, where $w_i := x_i / \sum_{i=1}^{N} x_i$. This means that proportion of weekly income spent on food for a population, $\sum_{i=1}^{N} y_i / \sum_{i=1}^{N} x_i$, is also a weighted average of proportions of weekly income spent on food per household, where the weight is proportional to the weekly income.

The data set, comprised of the weekly expenditures on food and weekly incomes for twenty households, is presented in Table 2 and depicted in Figure 4(a). The posterior distributions of β and σ are presented in Figures 4(b) and (c). The posterior medians of β are 0.283, 0.306 and 0.319 with 95% highest posterior density (HPD) intervals of $(0.217, 0.348)$, $(0.243, 0.367)$ and $(0.240, 0.376)$ for the nonrobust, partially robust and robust models, respectively. As a result, the proportion of weekly income spent on food for this population is estimated at 0.319 (considering our robust model) with a 95% HPD interval of $(0.240, 0.376)$. The average weekly household expenditure on food of this population can also be estimated using the ratio estimator. Considering our robust model, it is estimated at $\hat{\mu}_y = \hat{\beta} \times \mu_x = 0.319 \times 210 = 66.99$ (considering an average weekly household income of 210 for this population) with a 95% HPD interval of $(50.40, 78.96)$. The posterior medians of σ are 2.180, 2.031 and 1.634 with 95% HPD intervals of $(1.565, 3.016)$, $(1.319, 2.958)$ and $(0.962, 2.674)$, for the nonrobust, partially robust and robust models, respectively.

We observe the presence of two clear outliers: $(x_{17}, y_{17}) = (250.2, 6.1)$ (because of its extremely low y value) and $(x_{20}, y_{20}) = (696.4, 41.1)$ (because of its extremely high x value). In order to draw conclusions based on the bulk of the data and to evaluate the impact of outliers, we redo the analysis while excluding these two outliers. The results are presented in Figure 5. The estimates

Table 2 *Weekly expenditure on food* (y_i) *and weekly income* (x_i) *for household* i *in dollars,* $i = 1, \ldots, 20$

y_i	31.7	68.4	54.4	53.5	78.4	66.4	64.1	44.6	99.0	53.3
x_i	102.9	144.9	155.8	176.5	177.4	182.2	197.9	199.2	211.3	215.9
y_i	67.3	68.6	63.0	100.6	82.2	113.4	6.1	76.6	92.7	41.1
x_i	216.0	216.7	220.3	222.8	229.0	250.0	250.2	275.4	342.4	696.4

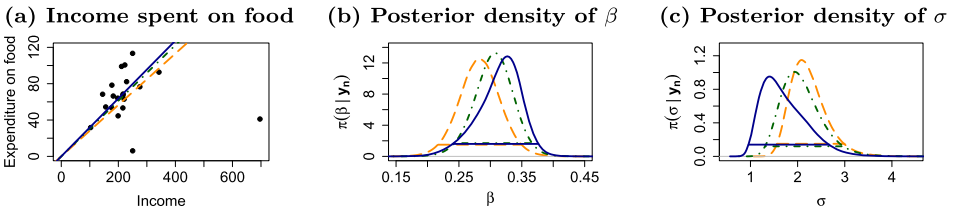

Figure 4 *Expenditure on food as a function of the income with an estimation of the expenditure on food* $\hat{\beta}x_i$ *based on the posterior median,* (b)–(c) *posterior densities of* β *and* σ *arising from the original data with 95% HPD intervals (horizontal lines); for each graph, the orange dashed, green dot-dashed and blue solid lines are respectively related to the nonrobust, partially robust and robust models.*

arising from the three models are now similar. The posterior medians of β are 0.342, 0.339 and 0.343 with 95% HPD intervals of $(0.302, 0.382)$, $(0.298, 0.380)$ and $(0.303, 0.382)$ for the nonrobust, partially robust and robust models, respectively. Therefore, the proportion of weekly income spent on food for this population is estimated at 0.343 (considering our robust model) with a 95% HPD interval of $(0.303, 0.382)$, based on the bulk of the data. Considering our robust model, the average weekly household expenditure on food is now estimated at $\hat{\mu}_y = \hat{\beta} \times \mu_x = 0.343 \times 210 = 72.03$ (considering an average weekly household income of 210 for this population) with a 95% HPD interval of $(63.63, 80.22)$, using the ratio estimator. The posterior medians of σ are 1.177, 1.268 and 1.190 with 95% HPD intervals of $(0.825, 1.656)$, $(0.850, 1.823)$ and $(0.854, 1.661)$, for the nonrobust, partially robust and robust models, respectively.

Based on the original data set, the inference arising from our robust model is the one that best reflects the behaviour of the bulk of the data, compared to the inferences arising from the nonrobust and partially robust models. Our robust model therefore succeeds in limiting the influence of outliers in order to obtain conclusions consistent with the majority of the observations.

Note that an outlier with an extreme x value, as $(x_{20}, y_{20}) = (696.4, 41.1)$, can be viewed as an observation with a fixed x value and an extreme y value (in this case, as an observation with a fixed x value of 696.4 and an extremely low y value of 41.1, compared to the trend emerging from the bulk of the data). This explains

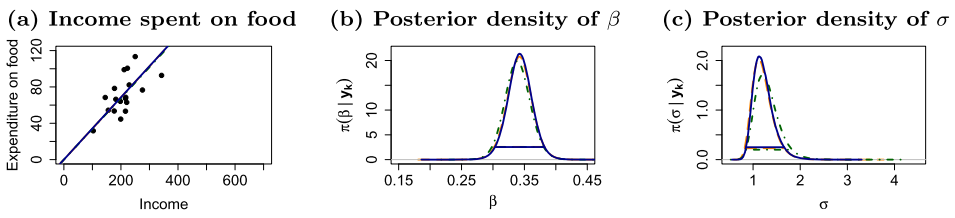

Figure 5 *Expenditure on food as a function of the income with an estimation of the expenditure on food* $\hat{\beta}x_i$ *based on the posterior median, when the outliers are excluded*, (b)–(c) *posterior densities of* β *and* σ *arising from the data set excluding the outliers with* 95% *HPD intervals* (*horizontal lines*); *for each graph, the orange dashed, green dot-dashed and blue solid lines are respectively related to the nonrobust, partially robust and robust models.*

why our robust model produces robust inference in the presence of this type of outliers.

3.2 Simulation study

We now evaluate the accuracy of the estimates arising from our robust model in a context of finite population sampling. More precisely, the model $Y_i = \beta x_i + \varepsilon_i$ with $\varepsilon_i \mid \sigma \overset{\mathcal{D}}{\sim} 1/(\sigma x_i^{1/2}) f(\varepsilon_i/(\sigma x_i^{1/2}))$ and $x_i > 0$, $i = 1, \ldots, n$, is used to analyse the data, where f is assumed to be a LPTN density in our robust model. We consider two sets of parameters for the LPTN: $(\alpha, \phi) = (1.96, 4.08)$ as in Section 3.1, and $(\alpha, \phi) = (1.50, 2.18)$. Our model is compared with the same linear regression model, but where f is assumed to be a standard normal density in the nonrobust model, and where f is assumed to be a Student density with 10 degrees of freedom and a known scale parameter of 0.88 in the partially robust model, as in Section 3.1. We set $\pi(\beta, \sigma) \propto 1$ and we estimate β and σ using MAP estimation (which therefore corresponds to maximum likelihood estimation) for these three models. Given that the obtained estimates are the same as under the frequentist paradigm, we also compare with the *M*- and *S*-estimators.

We set $n = 20$ and $x_1, x_2, \ldots, x_{20} = 1, 2, \ldots, 20$. We simulate 1,000,000 data sets using values for β and σ arbitrarily set to 1 and 1.5, respectively, and we carry out this process for each of the three scenarios that we now describe. In the first one, f is a standard normal distribution; therefore, the probability to observe outliers is negligible. In the second scenario, f is a mixture of two normals where the first component is a standard normal distribution and the second has a mean of 0 and a variance of 10^2, with weights of 0.9 and 0.1, respectively. This last component can contaminate the data set by generating extreme values. In the third and last scenario, f is also a mixture of two normals, but the contamination is due to the second component's location. More precisely, the first component is again a standard normal, but the second has a mean of 10 and a variance of 1, with weights of 0.95 and 0.05, respectively.

Table 3 *MSE of the estimators of β under the three scenarios*

Assumptions on f	Scenarios		
	$100\%\mathcal{N}(0,1)$	$90\%\mathcal{N}(0,1)+10\%\mathcal{N}(0,10^2)$	$95\%\mathcal{N}(0,1)+5\%\mathcal{N}(10,1)$
Standard normal	0.011	0.117	0.110
Student (10 d.f.)	0.011	0.027	0.033
LPTN			
with $\alpha = 1.96$ and $\phi = 4.08$	0.011	0.020	0.018
with $\alpha = 1.50$ and $\phi = 2.18$	0.013	0.016	0.013
M-estimator	0.011	0.017	0.017
S-estimator	0.029	0.027	0.027

Table 4 *MSE of the estimators of σ under the three scenarios*

Assumptions on f	Scenarios		
	$100\%\mathcal{N}(0,1)$	$90\%\mathcal{N}(0,1)+10\%\mathcal{N}(0,10^2)$	$95\%\mathcal{N}(0,1)+5\%\mathcal{N}(10,1)$
Standard normal	0.06	12.98	5.03
Student (10 d.f.)	0.06	4.02	1.99
LPTN			
with $\alpha = 1.96$ and $\phi = 4.08$	0.07	0.60	0.22
with $\alpha = 1.50$ and $\phi = 2.18$	0.09	0.20	0.11
M-estimator	0.14	0.24	0.21
S-estimator	0.11	0.23	0.15

Within each simulation scenario, we evaluate the performance of each model and estimator using sample mean square errors (MSE), based on the true values $\beta = 1$ and $\sigma = 1.5$. The results are presented in Tables 3 and 4.

If we first compare the models that we considered in Section 3.1 (the models with the normal, Student and LPTN with $\alpha = 1.96$ and $\phi = 4.08$ assumptions), we observe that they have *almost* identical performances for both the estimation of β and σ, when there are no outliers (the $100\%\ \mathcal{N}(0,1)$ scenario). This was expected given that the three related densities are very similar, especially on the interval $[-1.96, 1.96]$ where they all have 95% of their mass. They however differ in the thickness of their tails, and this feature plays a major role when the sample contains outliers, which is frequently the case for the two other scenarios. As expected, the presence of outlying observations has a major impact on the estimations when the traditional standard normal assumption is used. For the model with the Student distribution assumption, outliers influence the estimation of σ significantly, while having a lesser effect on $\hat{\beta}$, which reflects the partial robustness of this approach. The impact on the estimation of both β and σ is limited for our robust alternative, as suggested by the theoretical results.

Our robust model with $\alpha = 1.96$ and $\phi = 4.08$ performs better than the frequentist competitors regarding the estimation of σ in the absence of outliers. The latter however produce more accurate estimates in the probable presence of outliers. There is a trade-off between the extent to which a model (or a loss function for the frequentist competitors) matches the traditional normal one, and the level of robustness it features. We clearly observe this by decreasing the value for α of the LPTN to 1.5, which leads to a density that matches that of the normal on $[-1.5, 1.5]$ (instead of on $[-1.96, 1.96]$), but has heavier tails ($\phi = 2.18$).

4 Conclusion

In this paper, we have provided a simple Bayesian approach to robustly estimate both parameters β and σ of a simple linear regression through the origin, in which the variance of the error term can depend on the explanatory variable. It leads to robust estimators of finite population means and ratios. The approach is to replace the traditional normal assumption on the error term by a super heavy-tailed distribution assumption. In particular, we considered log-regularly varying distributions. Whole robustness is attained provided that both the negative and positive slope outliers are fewer than the nonoutlying observations, that is, $m < k$ and $p < k$, as stated in Theorem 1.

The theoretical results have been illustrated in Section 3 through typical real-life situations in which ratio estimation is used, and a simulation study. All the analyses leading to robust inference have been done using the log-Pareto-tailed standard normal (LPTN) density given in (2). Our model has been compared with the nonrobust (with the normal assumption) and partially robust (with the Student distribution assumption) models. The conclusion is: our model performs as well as the nonrobust and the partially robust models in absence of outliers, in addition to being completely robust. Therefore, our recommendation is to assume that the error has the density given in (2) and obtain adequate results, regardless of whether there are outliers, by computing estimates as usual from the posterior distribution.

Acknowledgments

The authors acknowledge support from the NSERC (Natural Sciences and Engineering Research Council of Canada), the FRQNT (Le Fonds de recherche du Québec—Nature et technologies) and the SOA (Society of Actuaries). They also would like to thank the anonymous referees for their very helpful comments.

Supplementary Material

Supplement to "Bayesian robustness to outliers in linear regression and ratio estimation" (DOI: 10.1214/17-BJPS385SUPP; .pdf). In the supplementary

material (Desgagné and Gagnon (2019)), you will find the proofs of Proposition 1 and Theorem 1 from our paper, and the R functions that were used for the computations.

References

Andrade, J. A. A. and O'Hagan, A. (2011). Bayesian robustness modelling of location and scale parameters. *Scandinavian Journal of Statistics* **38**, 691–711. MR2859745

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

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association* **81**, 1063–1069. MR0867633

Desgagné, A. (2013). Full robustness in Bayesian modelling of a scale parameter. *Bayesian Analysis* **8**, 187–220. MR3036259

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

Desgagné, A. and Gagnon, P. (2019). Supplement to "Bayesian robustness to outliers in linear regression and ratio estimation". DOI:10.1214/17-BJPS385SUPP.

Gwet, J.-P. and Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association* **87**, 1174–1182.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* **1** 799–821. MR0356373

Mallows, C. L. (1975). On some topics in robustness. Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

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

Peña, D., Zamar, R. and Yan, G. (2009). Bayesian likelihood robustness in linear models. *Journal of Statistical Planning and Inference* **139**, 2196–2207.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*, 256–272. Berlin: Springer.

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

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

Département de mathématiques et de statistique
Université de Montréal
C.P. 6128, Succursale Centre-ville
Montréal, Québec, H3C 3J7
Canada
E-mail: gagnonp@dms.umontreal.ca