Chilean Journal of Statistics
Vol. 1, No. 1, April 2010, 35–58

Special Issue "In Memory of Pilar Loreto Iglesias Zuazola"
Research Paper

Some thoughts on the Bayesian robustness of location-scale models

Manuel Mendoza[1,*] and Eduardo Gutiérrez-Peña[2]

[1]Department of Statistics, Instituto Tecnológico Autónomo de México, México D.F., México,
[2]Department of Probability and Statistics, Universidad Nacional Autónoma de México, México D.F., México

(Received: 02 February 2010 · Accepted in final form: 31 March 2010)

Abstract

In this paper we review a number of results, widely discussed in the literature, concerning the Bayesian robustness of location-scale models. We underline some specific aspects which, in our opinion, deserve more attention, and illustrate our remarks by means of a simple case study.

Keywords: Compatible priors · Conjugate priors · Elliptical models · Improper priors · Insensitivity · Marginal equivalence.

Mathematics Subject Classification: Primary 62F15 · Secondary 62G35.

1. Introduction

For more than 40 years, the idea of robustness has been explored within the Bayesian framework. Many authors have contributed to the field using different approaches. A well defined line of research, however, can be identified which applies to location-scale (regression) models. There, robustness is investigated in terms of the effect which can be observed when the sampling distribution is modified. To this end, the corresponding prior is often taken as improper. Results are quite general and interesting. In this paper, we review several contributions of this type, underlining some specific aspects which, in our opinion, deserve more attention. We then illustrate our remarks through a simple case study.

The outline of the paper is as follows. In the next section we discuss the statistical concept of robustness, both from a frequentist and a Bayesian point of view. In Section 3 we review the standard linear regression model as well as a seminal paper by Zellner (1976) concerning the robustness of marginal inferences on the regression coefficients when the assumption of normality is replaced with the assumption that the errors follow a Student-t distribution. We then discuss several generalizations of this result in Sections 4 and 5. Section 6 summarizes all of these results and attempts to put them in perspective. Section 7 presents a simple case study which we feel clarifies some aspects of Bayesian robustness in this setting. Finally, Section 8 contains some concluding remarks. A brief account of elliptical models is provided in Appendix.

[*] Corresponding author. Email: mendoza@itam.mx

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

2. Robustness

2.1 The general idea

Since the introduction of robustness as a statistical concept by Box (1953), a large amount of contributions dealing with different aspects of the subject have appeared, ranging from general, theoretical, investigations to specific robustness studies for particular applications.

Many other authors have contributed to the development of robustness as a vigorous research area in statistics. See, e.g., Andrews et al. (1972), Hampel et al. (1986), Hoaglin et al. (1983), Huber (1981), Huber (2009), Morgenthaler and Tukey (1991), Tukey (1960), Tukey (1977), and the references therein.

When a statistical procedure is used to describe a phenomenon, the final inference (optimal according to some criteria) is based both on a data set and on a number of assumptions which typically take the form of a specific model. Under such circumstances, the purpose of a robustness study is to:

(i) investigate if the procedure leads to results which are close to the optimal even when the data is atypical in some sense, or the assumptions are not met to a reasonable degree;
(ii) develop alternative (robust) procedures with the desired properties.

Several authors refer explicitly to these goals. For example, Morgenthaler (2007) says: "Robustness of statistical methods in the sense of insensitivity to grossly wrong measurements is probably as old as the experimental science". Huber (2002), speaking about lack of robustness, writes: "...the extreme sensitivity of some conventional statistical procedures to seemingly minor deviations from assumptions". More recently, Huber (2009) has elaborated on this:

> "Any statistical procedure should possess the following desirable features:
>
> - Efficiency: It should have a reasonably good (optimal or nearly optimal) efficiency at the assumed model.
>
> - Stability: It should be robust in the sense that small deviations from the model assumptions should impair the performance only slightly...
>
> - Breakdown: Somewhat larger deviations from the model should not cause a catastrophe."

So, in general terms, robustness seeks protection against (a few) large deviations in the data and relatively small deviations from assumptions. It should be noted, however, that lack of robustness is not always an undesirable property. In the classical example of a location parameter (mean versus median), the sample mean is not robust against outliers and, even so, there are some instances where it is preferable to the sample median as a summary of the data precisely because of its sensitivity to extreme observations.

On a more technical note, robustness, specially with respect to the sampling model, can be formulated as a continuity condition (given a suitable topology) of an operator relating a certain type of inferential result to a class of distributions (Huber, 2009). In any case, it is interesting to note that, from this point of view, an operator which is constant over a number of qualitatively different classes of distributions would be classified as robust even though it might be better called completely insensitive or invariant. Another issue is that of deciding when and where this type of insensitivity (invariance) can be regarded as a desirable property. We shall return to this idea in the following sections.

2.2 Bayesian robustness

From a Bayesian point of view, all the information regarding a quantity of interest is described through a probability distribution, which, together with a loss function, is the basis for any inference or decision. Robust Bayesian analysis is thus concerned with the sensitivity of the results to the inputs (the probability distribution, the loss function, or some combination thereof).

Specifically, in the usual setting of Bayesian parametric inference, there is a vector of observations $\boldsymbol{y} \in \mathbb{R}^n$ with sampling distribution $p(\boldsymbol{y}|\theta)$ and a prior distribution $p(\theta)$. The required posterior distribution $p(\theta|\boldsymbol{y})$ is obtained, via Bayes' theorem, from the sampling model (likelihood function) and the prior distribution as $p(\theta|\boldsymbol{y}) = p(\theta)p(\boldsymbol{y}|\theta)/p(\boldsymbol{y})$.

Hence, Bayesian robustness, in this context, can be (and should be) evaluated with respect to the sampling model (as in the frequentist approach), the prior distribution, and the loss function. In fact, from a Bayesian perspective the model is $p(\boldsymbol{y}, \theta)$, given by

$$p(\boldsymbol{y}, \theta) = p(\boldsymbol{y}|\theta)p(\theta),$$

and robustness of the posterior distribution should be assessed with respect to this joint distribution. It may happen, and we shall come back to this later, that two different sampling models produce the same joint distribution if combined with two suitable (different) priors. Nevertheless, for the sake of simplicity, most analyses proceed by fixing one of the components and exploring the robustness with respect to the other. More generally, robustness with respect to one component (prior/likelihood/loss) must be understood as conditional on the choice of the others.

As pointed out above, the basic idea is to evaluate the change in the output (e.g. the posterior distribution or a point estimate) associated with a change in the inputs (prior/likelihood/loss). Thus, a measure of variation in the output space is required. Different choices for this measure lead to different analyses. On the other hand, one can envisage several different approaches. The simplest one is an informal analysis where only a reduced (usually finite) set of alternative inputs is considered. A more general (global) approach considers the class of all alternative inputs compatible with the available information, and then computes the range of the output as the inputs vary over that class. Yet another (local) approach focuses instead on the output's rate of change. In addition, the class of alternative inputs can be regarded either as a parametric or a nonparametric family. The combination of all these possibilities has led to a rich field of research to which a large number of authors have contributed, mostly in the 90s. Excellent accounts of this work can be found, for example, in Berger (1990), Berger (1994) and, more recently, Ríos-Insua and Ruggeri (2000). Apart from the richness of ideas and the diversity of approaches, this latter volume reflects the relative emphasis that researchers have put on each of the components prior/likelihood/loss. Most of the contributions therein deal with robustness with respect to the prior. Only a few authors explore robustness with respect to the loss function and, remarkably, only one contribution deals with robustness with respect to the likelihood.

In the remainder of this paper we shall be mainly concerned with global robustness regarding the likelihood. One of the first such Bayesian robustness studies is due to Zellner (1976), who examined the problem from both the frequentist and Bayesian approaches for the linear regression model. We shall discuss his paper and some generalizations in the next section.

3. Linear Regression

Zellner (1976) explores the impact of replacing the usual normal distribution with a multivariate Student-t distribution on the inferential results from the linear multiple regression model. In order to describe his results more easily and set up the notation, let us briefly review the analysis of the normal model under the usual assumption.

3.1 The standard model

The standard, normal linear regression model is defined as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n),$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{X} = [x_{ij}]$ is a full-rank $(n \times p)$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. Here $\boldsymbol{I}_n$ denotes the identity matrix of order n. Then

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) = \mathrm{N}_n(\boldsymbol{y} \mid \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n), \tag{1}$$

and note that

$$\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2) = \boldsymbol{X\beta} \quad \text{and} \quad \mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2) = \sigma^2 \boldsymbol{I}_n.$$

3.2 Frequentist inference

Recall that the maximum likelihood estimators for this model are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

and that the likelihood ratio tests for the usual hypotheses concerning $\boldsymbol{\beta}$ and σ^2 result in t and χ^2 statistics, respectively. For the purpose of investigating robustness, Zellner (1976) assumes a Student-t distribution for the vector $\boldsymbol{\epsilon}$, so that

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) = St_n(\boldsymbol{y} \mid \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n, \nu_0). \tag{2}$$

Therefore, $\boldsymbol{Y}$ follows a multivariate Student-t distribution with location parameter $\boldsymbol{X\beta}$, scale matrix $\sigma^2 \boldsymbol{I}_n$ and ν_0 degrees of freedom. It must be noted that, instead of assuming independent univariate Student-t errors, Zellner (1976) proposes a multivariate Student-t distribution whose components are uncorrelated but no longer independent. Under these circumstances,

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X\beta} \quad (\nu_0 > 1) \quad \text{and} \quad \mathrm{Var}(\boldsymbol{y}) = \{\nu_0 \sigma^2 / (\nu_0 - 2)\} \boldsymbol{I}_n \quad (\nu_0 > 2).$$

Making use of the explicit form of the Student-t density function, Zellner (1976) proves that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are, again, the maximum likelihood estimators of the respective parameters. Here, it must be noted that σ^2 is no longer the variance of the observations but a multiple of that quantity (assuming $\nu_0 > 2$). Moreover, Zellner (1976) shows that the usual pivotal quantity for making inferences regarding σ^2 no longer has a χ^2 but an F distribution. Using the representation of the Student-t as a scale-mixture of normal distributions, where the mixing distribution is an inverted gamma, Zellner (1976) also proves that, in the case of $\boldsymbol{\beta}$, the t and F statistics remain unchanged so that inferences can be carried out exactly as in the normal case. Therefore, he concludes that the frequentist results are robust (with the exception of the pivotal quantity for σ^2) if the sampling distribution is modified from a normal to a Student-t model.

3.3 Bayesian inference

For the sake of comparison, let us now briefly describe the usual Bayesian results under the normal sampling model and two commonly used prior distributions.

Conjugate prior Let $\phi = 1/\sigma^2$. The standard conjugate prior for Equation (1) is the normal-gamma distribution, whose density is denoted by

$$\begin{aligned} p(\boldsymbol{\beta}, \phi) &= p(\boldsymbol{\beta} \,|\, \phi)\, p(\phi) \\ &= \mathrm{N}_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_0, \phi^{-1}\boldsymbol{B}_0^{-1})\, \mathrm{Ga}(\phi \,|\, d_0/2, a_0/2). \end{aligned} \tag{3}$$

This distribution is proper provided that $a_0 > 0$, $d_0 > 0$ and $\boldsymbol{B}_0$ is positive-definite.

Posterior distribution The posterior distribution for the parameters of Equation (1) corresponding to the conjugate prior, Equation (3) is

$$p(\boldsymbol{\beta}, \phi \,|\, \boldsymbol{y}) = \mathrm{N}_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_1, \phi^{-1}\boldsymbol{B}_1^{-1})\, \mathrm{Ga}(\phi \,|\, d_1/2, a_1/2),$$

where

$$\begin{aligned} \boldsymbol{b}_1 &= (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{B}_0)^{-1}(\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{B}_0 \boldsymbol{b}_0) \\ \boldsymbol{B}_1 &= \boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{B}_0 \\ a_1 &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_1)^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_1) + (\boldsymbol{b}_1 - \boldsymbol{b}_0)^\top \boldsymbol{B}_0(\boldsymbol{b}_1 - \boldsymbol{b}_0) + a_0 \\ d_1 &= n + d_0. \end{aligned}$$

Hence, the marginal posterior for ϕ is

$$p(\phi \,|\, \boldsymbol{y}) = \mathrm{Ga}(\phi \,|\, d_1/2, a_1/2).$$

On the other hand, the marginal posterior for $\boldsymbol{\beta}$ is given by

$$\begin{aligned} p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = &\frac{\Gamma((d_1 + p)/2)}{\Gamma(d_1/2)\,\pi^{p/2}} \left| \frac{1}{a_1}\boldsymbol{B}_1 \right|^{1/2} \\ &\times \left\{ 1 + \frac{1}{a_1}(\boldsymbol{\beta} - \boldsymbol{b}_1)^\top \boldsymbol{B}_1(\boldsymbol{\beta} - \boldsymbol{b}_1) \right\}^{-(d_1+p)/2}. \end{aligned}$$

That is, a Student-t distribution with d_1 degrees of freedom. We denote this density by

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = St_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_1, \boldsymbol{T}_1^{-1}, d_1),$$

where $\boldsymbol{T}_1 = \left(\frac{d_1}{a_1}\right)\boldsymbol{B}_1$. Finally, recall that

$$\mathrm{E}(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \boldsymbol{b}_1 \text{ (if } d_1 > 1) \quad \text{and} \quad \mathrm{Var}(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = \frac{a_1}{d_1 - 2}\boldsymbol{B}_1^{-1} \text{ (if } d_1 > 2).$$

Reference prior and posterior In strict justice, it must be realized that, at the time Zellner (1976) was working on his paper, the idea of reference priors as we know it today was not all that clear. Jeffrey's rule was certainly already in use but his priors were most commonly referred to as noninformative. In fact, Zellner (1976) refers to a diffuse prior for $(\boldsymbol{\beta}, \phi)$ given by

$$\pi(\boldsymbol{\beta}, \phi) \propto \phi^{-1}, \tag{4}$$

This prior is improper, corresponds to the Jeffrey's rule assuming independence between $\boldsymbol{\beta}$ and ϕ, and can be formally seen as a limiting case of the conjugate prior, Equation (3), as $a_0 \to 0$, $d_0 \to -p$ and $\boldsymbol{B}_0 \to \boldsymbol{O}$. Consequently, the reference posterior distribution can also be seen as a limiting case of the corresponding conjugate posterior distribution and is characterized by

$$\begin{aligned} \boldsymbol{b}_1 &= \hat{\boldsymbol{\beta}} \\ \boldsymbol{B}_1 &= \boldsymbol{X}^\top \boldsymbol{X} \\ a_1 &= (n-p)\,\hat{\sigma}^2 \\ d_1 &= n-p. \end{aligned}$$

Hence,

$$\pi(\phi \,|\, \boldsymbol{y}) = \mathrm{Ga}(\phi \,|\, (n-p)/2, (n-p)\hat{\sigma}^2/2)$$

and

$$\pi(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = St_p(\boldsymbol{\beta} \,|\, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, n-p).$$

Robust analysis Instead of the usual normal assumption, the Student-t model (2), is adopted as the sampling distribution and then, for the fixed prior, Equation (4), Zellner (1976) shows that the marginal posterior distribution for $\boldsymbol{\beta}$ is exactly the same as in the normal case (a Student-t), whereas the marginal posterior for σ^2 is an F distribution. The overall conclusion is that inferences regarding the regression coefficients are robust (under both the frequentist and Bayesian approaches) when the sampling distribution is taken to be a multivariate Student-t instead of normal. This result has been generalized in several ways. In the next sections we shall review some of the most interesting cases.

4. Generalizations

4.1 Nonlinear regression: other mixing distributions

Osiewalski (1991) generalizes Zellner's result, within an entirely Bayesian framework, by dealing with a possibly nonlinear regression model of the form

$$\boldsymbol{Y} = h(\boldsymbol{X};\boldsymbol{\beta}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} = \psi(Z) \times \boldsymbol{U}, \quad \boldsymbol{U}|\boldsymbol{\beta},\eta,\phi \sim \mathrm{N}_n(\boldsymbol{0}, \phi^{-1}\boldsymbol{V}(\eta)),$$

where $\boldsymbol{V}(\eta)$ is a known matrix function of the unknown parameter η, $\psi(\cdot)$ is a positive function, and Z is a continuous positive random variable with density $p(z|\boldsymbol{\beta},\eta,\phi)$. Thus, the vector $\boldsymbol{\epsilon}$ has a distribution given by a general scale mixture of multivariate normal distributions and the sampling model is not necessarily a multivariate Student-t but a special type of elliptical distribution (see Appendix for a definition). Under these conditions, Osiewalski (1991) proves that for the rather general family of prior distributions

$$p(z,\phi,\boldsymbol{\beta},\boldsymbol{\eta}) = p(z,\phi|\boldsymbol{\beta},\boldsymbol{\eta})\,p(\boldsymbol{\beta},\boldsymbol{\eta}),$$

where

$$p(z,\phi|\boldsymbol{\beta},\boldsymbol{\eta}) = \frac{c}{\phi}\,p(z|\boldsymbol{\beta},\boldsymbol{\eta}),$$

the marginal posterior distribution of $\boldsymbol{\beta}$ is the same as for the case where the error is multivariate normal. An analogous result is obtained for the predictive distribution. The most noticeable condition here is that of an improper prior density which in terms of ϕ is proportional to ϕ^{-1}.

4.2 Nonlinear regression: general elliptical distributions

In Osiewalski and Steel (1993a), the previous results are generalized to the case of an arbitrary multivariate elliptical distribution for the vector of errors. Again, the marginal posterior distribution for the coefficients as well as the prior predictive distribution are the same as those for the normal case. Once more, the structure includes a common precision parameter ϕ whose prior density must be proportional to ϕ^{-1}. Further, the authors prove the same result for a sampling distribution defined as a finite mixture of elliptical models where the elliptical components have the same parameters but differ in both the covariance and nonlinear structures.

4.3 Marginal equivalence

Osiewalski and Steel (1993b) introduce the idea of marginal equivalence for Bayesian modeling. Given two parametric families of sampling models,

$$P = \{p(\boldsymbol{y}|\boldsymbol{\beta}, \delta)\colon \boldsymbol{\beta} \in \boldsymbol{B}, \delta \in \boldsymbol{D}\} \quad \text{and} \quad P_* = \{p_*(\boldsymbol{y}|\boldsymbol{\beta}, \phi)\colon \boldsymbol{\beta} \in \boldsymbol{B}, \phi \in \boldsymbol{F}\},$$

let $p(\boldsymbol{\beta}, \delta) = p(\delta|\boldsymbol{\beta})p(\boldsymbol{\beta})$ and $p_*(\boldsymbol{\beta}, \delta) = p_*(\delta|\boldsymbol{\beta})p_*(\boldsymbol{\beta})$ be the corresponding prior distributions. If $\boldsymbol{\beta}$ has the same interpretation in both families, so that $p(\boldsymbol{\beta}) = p_*(\boldsymbol{\beta})$, the Bayesian models $p(\boldsymbol{y}, \boldsymbol{\beta}, \delta)$ and $p_*(\boldsymbol{y}, \boldsymbol{\beta}, \phi)$ are said to be marginally equivalent if $p(\boldsymbol{y}, \boldsymbol{\beta}) = p_*(\boldsymbol{y}, \boldsymbol{\beta})$. It is clear that two marginally equivalent models produce the same posterior distribution for the parameter (of interest) $\boldsymbol{\beta}$ and the same predictive distribution for $\boldsymbol{Y}$ if the marginal model $p(\boldsymbol{y}, \boldsymbol{\beta})$ is proper. Under these conditions, the authors fix, as the family P_*, a nonlinear normal regression model where the covariance matrix is known up to a scalar precision factor. For this parameter, they choose a particular form of a gamma distribution as its conditional prior distribution given $\boldsymbol{\beta}$. As for the family P, they keep the same nonlinear regression structure and consider two different instances of sampling distributions, namely a Student-t model and a Pearson type II model. In the former case, they find that a beta conditional prior for the precision parameter yields marginal equivalence, whereas for the latter case the same results is obtained from an inverted beta.

An interesting idea in this paper is that a model with heavier tails than the normal, such as the Student-t, can produce the same results as the normal if, taking advantage of the fact that the scale factor is unknown in both models, these parameters are used to equate the variations of the observable quantity.

4.4 Nonlinear regression: a broader class of distributions

In Fernández et al. (1997), a general class of multivariate distributions is introduced by defining a correspondence between points $\boldsymbol{z}$ in $\mathbb{R}^n - \{\boldsymbol{0}\}$ and pairs $(\boldsymbol{w}, r)$ where $r > 0$ and $\boldsymbol{w}$ lies in some $(n-1)$-manifold $\mathcal{W}$ which can be thought of as a general version of the $(n-1)$-sphere. The idea is to produce a generalization of the class of spherical distributions. Thus, a random vector $\boldsymbol{Z}$ defined on $\mathbb{R}^n$ is first represented by a random vector $\boldsymbol{W}$ defined on $\mathcal{W}$ together with a positive random variable R which plays the role of a generalized radius. Within this framework, specific subclasses of distributions can then be defined if, for example, $\boldsymbol{W}$ and R are independent or the marginal distribution of either of them is fixed. As for the topic of Bayesian robustness, these authors deal with the situation where a collection $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_p}$ of vectors in $\mathrm{I\!R}^n$ is observed in accordance to the model

$$\boldsymbol{Y}_i = \boldsymbol{b}_i(\boldsymbol{\beta}) + \sigma_i \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, p,$$

where the $\sigma_i > 0$ are scale parameters and the $\boldsymbol{b}_i$ are location vectors parameterized in terms of a common vector $\boldsymbol{\beta}$ in $\mathrm{I\!R}^m$ $(m \leq n)$. They use the improper prior

$$p(\boldsymbol{\beta}, \sigma_1, \ldots, \sigma_p) \propto p(\boldsymbol{\beta}) \prod_{i=1}^{p} \sigma_i^{-1}$$

and prove that the joint distribution of $(\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_p}, \boldsymbol{\beta})$ is exactly the same as long as the sampling distribution is obtained by taking $\boldsymbol{\epsilon} = R\boldsymbol{W}$ for a fixed distribution on $\boldsymbol{W}$ and an arbitrary distribution on R. Provided that the resulting distributions are proper, $p(\boldsymbol{y_1}, \ldots, \boldsymbol{y_p})$ and $p(\boldsymbol{\beta}|\boldsymbol{y_1}, \ldots, \boldsymbol{y_p})$ are then invariant with respect to the sampling distribution.

This result generalizes the case of spherical distributions, although it may lead to improper posterior distributions.

4.5 Influence measures for elliptical distributions

Arellano-Valle et al. (2000) consider the usual linear regression structure as in Section 3.1 but based on an elliptical sampling distribution, namely

$$[\boldsymbol{y} \,|\, \boldsymbol{X}, \boldsymbol{\beta}, \phi] \sim \mathrm{EL}_n(\boldsymbol{X}\boldsymbol{\beta}, \phi^{-1}\boldsymbol{I}_n; h(\cdot)),$$

with density given by

$$\phi^{n/2} h^n(\phi(\boldsymbol{y} - \boldsymbol{\mu})^\top (\boldsymbol{y} - \boldsymbol{\mu})),$$

and where $h(\cdot)$ is the generator of the class (see Appendix). For this model, they use the improper prior, Equation (4), and hence recover the corresponding invariance (robustness) results. On this basis, the authors explore a number of influence measures which have been previously considered for the normal case. As expected, under these conditions they show that any influence measure based on the posterior distribution of $\boldsymbol{\beta}$ is invariant with respect to the choice of the generator $h(\cdot)$. By contrast, influence measure concerning the scale parameter ϕ do depend upon $h(\cdot)$. More importantly, the authors suggest that monitoring the behavior of the influence measures based on the posterior distribution of the scale parameter might be useful to compare and choose among alternative generators.

In another paper, Arellano-Valle et al. (2003) propose a family of conjugate prior distributions which reproduce, to some extent, the invariance property of the improper prior $\pi(\boldsymbol{\beta}, \phi) \propto \phi^{-1}$.

Using results on marginal and conditional distributions for subvectors of a random vector with a spherical distribution, these conjugate priors define $p(\phi \,|\, h)$ as a radial distribution which combines with the density of the elliptical distribution for the regression model. As for $p(\boldsymbol{\beta} \,|\, \phi, h)$, the distribution is, in principle, arbitrary although some conditions are imposed to recover the invariance property. Under these conditions, they show that the invariance obtains whenever $p(\boldsymbol{\beta} \,|\, h)$ does not depend on h. Moreover, if this density is constant, then the marginal posterior distribution of $\boldsymbol{\beta}$ is a Student-t, just as in the normal model with a diffuse prior.

The idea of achieving invariance under a proper prior is clearly related to, and generalizes, the concepts of marginal equivalence and semiconjugate priors as discussed by Osiewalski and Steel (1993b). Arellano-Valle et al. (2003) go back to the diffuse prior case and explore, under these conditions, several known procedures for model comparison. They show that such procedures are useless for comparing models based on different generators, precisely due to the fact that some key expressions do not depend on the generator at all.

More recently, Arellano-Valle et al. (2006) go one step further along this line of research and define a class of models whose linear regression structure is embedded in a spherical sampling distribution where three supposedly known hyperparameters are included which, in particular, modify the role of ϕ as a scale parameter. The corresponding likelihood is then combined with a prior distribution $p(\boldsymbol{\beta}, \phi)$ chosen from a family such that $\boldsymbol{\beta}$ and ϕ are independent a priori, and where $p(\boldsymbol{\beta})$ is arbitrary and $p(\phi)$ is a squared radial distribution which shares the same generator and a specific hyperparameter with the likelihood. The authors then provide analytic expressions for the posterior distribution of $\boldsymbol{\beta}$ and ϕ, and show that for certain choices of the newly introduced hyperparameters, and if $p(\boldsymbol{\beta})$ is constant, the posterior distribution of $\boldsymbol{\beta}$ does not depend on the generator.

5. General Location-Scale Models

5.1 A first generalization

One of the first Bayesian robustness studies for location-scale families is that of Box and Tiao (1962), who introduce the family of exponential power distributions, with densities of the form

$$p(y \,|\, \mu, \phi, \gamma) \propto \phi \exp\left\{-\frac{1}{2}\left[\phi(y-\mu)\right]^{2/(1+\gamma)}\right\},$$

as a generalization of the normal distribution. Here, apart from the location parameter μ and the scale parameter ϕ, Box and Tiao (1962) introduce the shape parameter γ which, for $\gamma = 0$, reproduces the normal model. It is interesting to note that, since the normal is a particular instance of the spherical distribution, the exponential power can be regarded as the univariate version of an l_q spherical distribution (discussed below).

Box and Tiao (1962) analyze the robustness of the posterior distribution of μ with respect to the value of γ when the diffuse prior $\pi(\mu, \phi) \propto \phi^{-1}$ is used. They conclude that inferences regarding μ are not robust against changes in the value of γ. Also, these authors extend their analysis to linear and nonlinear regression models and reach similar conclusions.

More than thirty years later, Osiewalski and Steel (1993c) explore a rather different approach and introduce the l_q spherical models which generalize the usual spherical family by replacing the l_2 norm with the more general norm

$$v_q(\boldsymbol{a}) = \left(\sum_{i=1}^{n} a_i^q\right)^{1/q}.$$

Thus, a random vector $\boldsymbol{Y} \in \mathbb{R}^n$ is said to have an l_q spherical distribution with location parameter $\boldsymbol{\mu} \in \mathbb{R}^n$ and scale parameter $\phi \in \mathbb{R}_+$ if

$$p(\boldsymbol{y} \,|\, \boldsymbol{\mu}, \phi) = \phi^n g[v_q(\phi(\boldsymbol{y} - \boldsymbol{\mu}))],$$

where the generator function $g(\cdot)$ is such that $p(\boldsymbol{y} \,|\, \boldsymbol{\mu}, \phi)$ is proper for all $\boldsymbol{\mu} \in \mathbb{R}^n$ and

$\phi \in \mathbb{R}_+$. For this model they prove that, if

$$p(\boldsymbol{\mu}, \phi) \propto \phi^{-1} p(\boldsymbol{\mu}), \tag{5}$$

a priori, then $p(\boldsymbol{\mu} \,|\, \boldsymbol{y})$, when proper, depends on q but does not depend on $g(\cdot)$. Hence, given q, inferences about $\boldsymbol{\mu}$ are robust (invariant) with respect to the generator. It is interesting to note that, in this general setting, what the result by Box and Tiao (1962) says is that, for the (fixed) generator they use, inferences are not robust against changes in the form of the norm.

5.2 An enlarged spherical family

Fernández et al. (1994) show that the class of spherical distributions can be enlarged to include any continuous multivariate distribution. First recall that an ordinary spherical distribution for a vector $\boldsymbol{Y}$ defined on $\mathbb{R}^n$ can be interpreted as the product of a uniform distribution over the unit sphere $\mathcal{S}^{n-1}$ and a distribution for the Euclidean norm of the random vector (the radius). The generalization is achieved by weakening the independence condition and allowing distributions other than the uniform over the unit sphere $\mathcal{S}^{n-1}$. On this basis, they consider a vector $\boldsymbol{Y} \in \mathbb{R}^n$ such that $\boldsymbol{Y} = \boldsymbol{\mu} + \phi^{-1}\boldsymbol{Z}$ where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\phi > 0$ are fixed location and scale parameters, whereas $\boldsymbol{Z}$ is a random vector with a proper density $f(\cdot)$. Thus,

$$p(\boldsymbol{y} \,|\, \boldsymbol{\mu}, \phi) = \phi^n f(\phi(\boldsymbol{y} - \boldsymbol{\mu})),$$

defines a general location-scale family. In this setting, the authors prove, after decomposing $\boldsymbol{Z}$ into its projection onto $\mathcal{S}^{n-1}$ and its norm (radius), that a prior of the form of Equation (5) leads to a posterior distribution $p(\boldsymbol{\mu} \,|\, \boldsymbol{y})$ which, when proper, does not depend on the conditional distribution of the radius given the direction on $\mathcal{S}^{n-1}$. Therefore, this result states that, for any location-scale family, if the prior is given by Equation (5), then inferences concerning the location parameter are robust (invariant) with respect to the class of distributions for $\boldsymbol{Z}$ which share the same distribution over the unit sphere $\mathcal{S}^{n-1}$.

5.3 The v-spherical family

In a follow-up paper, Fernández et al. (1995) propose another generalization of the family of spherical distributions. A random vector $\boldsymbol{Y} \in \mathbb{R}^n$ has a v-spherical distribution with location parameter $\boldsymbol{\mu} \in \mathbb{R}^n$ and scale parameter $\phi \in \mathbb{R}_+$ if

$$p(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \phi) = \phi^n g[v(\phi(\boldsymbol{y} - \boldsymbol{\mu}))], \tag{6}$$

where $v(\cdot)$ is a positive homogeneous function, $g(\cdot)$ is a nonnegative function, and both are such that Equation (6) is a proper density.

The properties of this family are explored and, in particular, it is shown that if the prior, Equation (5), is adopted, the posterior distribution of $\boldsymbol{\mu}$, when proper, does not depend on $g(\cdot)$. In other words, for a fixed $v(\cdot)$, the models in the class defined by all possible choices of $g(\cdot)$ are marginally equivalent if $p(\boldsymbol{\mu}, \phi) \propto \phi^{-1} p(\boldsymbol{\mu})$. Some particular examples of marginal equivalence are also provided for the case where the prior distribution is proper.

5.4 Skew distributions

In a related direction, Vidal et al. (2006) explore how the usual analysis of the normal model can be extended when two different generalizations are considered simultaneously. On the one hand, the symmetric normal distribution is replaced with an elliptical (also symmetric) distribution. On the other hand, the symmetry condition is relaxed by considering skew distributions as introduced by Azzalini (1985). In order to get analytic results, the authors deal with a specific subclass of skew-elliptical models, namely those which are representable as a scale mixture of skew-normal distributions. Their main result is that the L_1 distance between a representable elliptical distribution and a representable skew-elliptical distribution remains invariant and is equal to the L_1 distance between the corresponding normal and skew-normal models. The authors also quantify the effect of the skewness parameter (λ) on the posterior distribution of the location (μ) and scale (σ) parameters. They assume a proper normal-inverse-gamma prior for (μ, σ^2) (independent of λ) and prove that the posterior density can be written as the product of the posterior density corresponding to the symmetric model and a perturbation function which depends on λ and measures the sensitivity of posterior inferences to the degree of skewness. In this case, the posterior distribution (in particular, that of the location parameter μ) is not invariant to the choice of the skew parameter. Nevertheless, the authors show that, for large sample sizes, the perturbation function tends to one and hence posterior inferences are asymptotically robust with respect to the degree of skewness.

6. Discussion

Each of the papers reviewed in Sections 4 and 5 introduces some nice and clever ideas to investigate robustness. In the case of regression, the starting point is the normal linear model with homoscedastic, uncorrelated errors. This setting is generalized to a situation where the location parameter has a nonlinear structure and the covariance structure is rather general apart from a common scale parameter ϕ. More importantly, the normality assumption is abandoned in favor of increasingly wider classes of models (Student-t, mixtures of normals, spherical and elliptical models, and even more general models).

In all of these cases, however, a univariate scale parameter is introduced and robustness of the posterior distribution of the regression coefficients $\boldsymbol{\beta}$ and/or the posterior predictive distribution is assessed mainly with respect to the sampling distribution. As for the other component of the model (the prior distribution) we can distinguish two cases. The more prominent uses an improper prior $p(\boldsymbol{\theta}, \phi) \propto \phi^{-1} p(\boldsymbol{\theta})$, with $p(\boldsymbol{\theta})$ either proper or improper, where $\boldsymbol{\theta}$ includes $\boldsymbol{\beta}$ and possibly other parameters. This idea appears for the first time in Zellner (1976) and allows him to relate Bayesian robustness with its frequentist counterpart. There, even though the prior might not be unique, it always belongs to a family of diffuse priors. The other case is that of a proper prior for ϕ. To deal with this situation, the general strategy does not fix the prior in advance and then explore different likelihoods. Instead, the sampling distribution and the corresponding prior on ϕ (marginal or conditional on $\boldsymbol{\beta}$) are chosen simultaneously whereas the prior for $\boldsymbol{\beta}$ is rather arbitrary.

Several authors have argued in favor of letting the prior for ϕ depend on the specific likelihood. For example, Arellano-Valle et al. (2003) say "The dependence... is reasonable, since in the present context the interpretation of the scale parameter changes with the density generator".

So, we would like to emphasize that, in the regression setting, robustness is analyzed with respect to the likelihood while the prior for ϕ is either diffuse (improper) or, being proper, changes according to the chosen likelihood.

When dealing with the closely related case of location-scale families, the situation is similar –if more transparent– since the use of covariates is almost entirely excluded and then robustness analysis is more explicitly focused on the underlying family of distributions.

As mentioned above, the pioneering work by Box and Tiao (1962) shows that inferences about the location parameter lack robustness when the normal model is replaced with other members of the exponential power family and a diffuse prior is considered.

On the other hand, in a series of papers, Fernández et al. (1994, 1995, 1997) deal with a rather different setting. First, instead of a sample of univariate observations, and in close relation with the regression framework, repeated sampling implies a collection of vectors. More importantly, the basic structure

$$p(\boldsymbol{y} \mid \boldsymbol{\mu}, \phi) = \phi^n f(\phi\,(\boldsymbol{y} - \boldsymbol{\mu})),$$

where $f(\cdot)$ is a completely arbitrary (multivariate, continuous) distribution, is approached by taking f to be either in the class of ordinary spherical distributions or in a specific subclass of the l_q-spherical or v-spherical families. They also introduce a general representation of any multivariate continuous distribution in terms of a distribution (not necessarily uniform) over the unit sphere $\mathcal{S}^{n-1}$ and a distribution for the respective Euclidean norm (independent of the former). Using this representation, they induce various classes of distributions by fixing the distribution over $\mathcal{S}^{n-1}$. As a general result, these authors prove that for $f(\cdot)$ within a suitable class, inferences about $\boldsymbol{\mu}$ or $\boldsymbol{Y}$ are robust if the prior for ϕ is $\pi(\phi) \propto \phi^{-1}$. Also, more restrictive results are obtained for proper prior distributions on ϕ, and in such cases, the prior depends on the corresponding $f(\cdot)$ under consideration.

Summing up, we can basically identify two scenarios. Either the diffuse prior is adopted and wide robustness results obtained, or a proper prior is used and it is still possible to achieve robustness, although in a more limited fashion and always adjusting the prior to the specific choice of the sampling distribution.

We then see that, for the purpose of assessing the robustness of the inferences with respect to the sampling distribution, in the elicitation process (both in the regression and location-scale settings) the key issue is the specification of the prior distribution for the scale parameter. As a consequence, extra caution is strongly advised in that respect.

It is also worth recalling that the location $\boldsymbol{\mu}$ or the vector of coefficients $\boldsymbol{\beta}$ is usually regarded as a well-defined parameter in the sense that it has the same interpretation regardless of the chosen model in a given class, whereas the scale parameter is recognized as having an interpretation which depends on the specific model. We feel that this aspect of the problem deserves a more careful discussion if one is to fully understand the implications of the results discussed so far.

In particular, it must be stressed that one of the original ideas behind the study of robustness was that of using sampling distributions with heavier tails than the normal in order to accommodate extreme observations revealing an extraordinary level of variability in the data (Box and Tiao, 1962 and Zellner, 1976).

Under these circumstances, it cannot be ignored that the general interpretation of the scale parameter is precisely in terms of dispersion. Thus, in the normal case, ϕ^{-1} can be directly interpreted as the variance. However, when the class of sampling models is generalized as in the papers reviewed in the preceding sections, if the second moment exists then the variance is, in general, given by $c_f\phi^{-1}$, where c_f is a proportionality constant which depends on the specific sampling distribution under consideration. Moreover, for other distributions the moments might not even exist. At this point, it is worth recalling de Finetti (1974), who states: “attention should be focused on the predictive distribution concerning directly the quantities of interest rather than the ones concerned with the parameters that are but auxiliary ingredients”.

Incidentally, it must be recognized that something similar may happen with the location parameter as well, since it can only be interpreted as the mean if the first moment exists. Thus, and again following de Finetti (1974), it might be preferable to describe our knowledge in term of truly common parameters, in the sense that: (i) they always exist (as limits of observable functions of the data); (ii) they have the same interpretation across models. Related remarks on this issue can be found in Gustafson (2001) and, more recently, Copas and Eguchi (2010). See also Section 1.4.5.1 of Berger et al. (2000).

An obvious proposal satisfying (i) and (ii) are the quantiles. In the univariate location-scale problem, if $Y = \mu + \phi^{-1}Z$, where $p(z) = f(z)$, then $p(y \mid \mu, \phi) = \phi f(\phi(y - \mu))$. Denoting by $z_{(\alpha)}$ the α-quantile of $f(z)$, then $y_{(\alpha)} = \mu + \phi^{-1}\, z_{(\alpha)}$ under $p(y \mid \mu, \phi)$.

Therefore, if $\alpha \neq \gamma$ are fixed, the prior distribution of (μ, ϕ) corresponding to $f(\cdot)$ can be easily obtained as a transformation of the elicited prior for $(y_{(\alpha)}, y_{(\gamma)})$ which does not depend on any specific model f. Hence, $p(\mu, \phi)$ will depend on $f(\cdot)$ only through $(z_{(\alpha)}, z_{(\gamma)})$. Moreover, if $f_1(\cdot)$ and $f_2(\cdot)$ are two elements of the class being explored for robustness, it is clear that coherence requires that the unique prior $p(y_{(\alpha)}, y_{(\gamma)})$ induce two different priors $p_1(\mu, \phi)$ and $p_2(\mu, \phi)$ for $f_1(\cdot)$ and $f_2(\cdot)$ respectively, as long as $\big(z^{(1)}_{(\alpha)}, z^{(1)}_{(\gamma)}\big) \neq \big(z^{(2)}_{(\alpha)}, z^{(2)}_{(\gamma)}\big)$.

In brief, difficulties associated with the interpretation of ϕ (as well as μ) can be overcome if these parameters are expressed in terms of quantiles of observable quantities and the prior distribution is then elicited on these quantiles (see Dickey and Chen, 1985, for a related discussion).

In general, these analyses preclude the use of the same prior for (μ, ϕ) when moving across different models. As a consequence, extra care is required when using the same improper prior $\pi(\mu, \phi) \propto \phi^{-1}$. This is in accordance with the findings of Arellano-Valle et al. (2003) regarding model comparison. A similar remark applies to the work by Vidal et al. (2006) on skew-elliptical distributions, where the authors use the same proper prior for the location and scale parameters regardless of the specific skew sampling model. This is probably not the best strategy to analyze the robustness of that generalization.

In the next section, a case study in the context of the linear regression model will be presented in order to illustrate some of these ideas.

7. A Case Study

7.1 A simple elliptical linear model

Here we consider the linear model based on the generators $h^n(u; k)$ defined in Appendix. Specifically,

$$[\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2] \sim \mathrm{EL}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n; h(\cdot; k)).$$

In other words,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}_n(\boldsymbol{0}, \sigma^2 k \boldsymbol{I}_n).$$

This is, acknowledgedly, a very simple model. However, it does allow us to compute all the required densities (posterior and predictive) in closed form while still shedding some light on the subtleties of the robustness analysis for more general elliptical and location-scale models.

The estimators of $\boldsymbol{\beta}$ and σ^2 for this model are given by

$$\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{\sigma}^2_k = \frac{1}{(n-p)k} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 / k.$$

Recall that $\phi = 1/\sigma^2$. Thus we can write

$$f_k(\boldsymbol{y} \mid \boldsymbol{\beta}, \phi) = \mathrm{N}_n(\boldsymbol{y} \mid \boldsymbol{X\beta}, \phi^{-1} k \boldsymbol{I}_n), \tag{7}$$

so that

$$\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{\beta}, \phi) = \boldsymbol{X\beta} \quad \text{and} \quad \mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{\beta}, \phi) = \phi^{-1} k \boldsymbol{I}_n.$$

We now consider two related but different conjugate families for the general normal model (7).

7.1.1 Näive prior

Suppose we take, for all values of k, the prior distribution for parameters of the general normal model to be the same as the prior distribution used in Section 3.3 for the standard linear model, namely

$$p^N(\boldsymbol{\beta}, \phi) = \mathrm{N}_p(\boldsymbol{\beta} \mid \boldsymbol{b}_0, \phi^{-1} \boldsymbol{B}_0^{-1})\, \mathrm{Ga}(\phi \mid d_0/2, a_0/2). \tag{8}$$

We call this the näive prior because it does not take into account the fact that the interpretation of the scale parameter ϕ is not the same for different values of k. Specifically, for all $i = 1, \ldots, n$,

$$\phi = \frac{k}{\mathrm{Var}(Y_i)}.$$

The posterior distribution under the näive prior is given by

$$p^N(\boldsymbol{\beta}, \phi \mid \boldsymbol{y}) = \mathrm{N}_p(\boldsymbol{\beta} \mid \boldsymbol{b}_1^N, \phi^{-1} [\boldsymbol{B}_1^N]^{-1})\, \mathrm{Ga}(\phi \mid d_1^N/2, a_1^N/2),$$

where

$$\begin{aligned}
\boldsymbol{b}_1^N \equiv \boldsymbol{b}_1(k) \; &= [(1/k)\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{B}_0]^{-1} [(1/k)\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{B}_0 \boldsymbol{b}_0] \\
\boldsymbol{B}_1^N \equiv \boldsymbol{B}_1^N(k) &= (1/k)\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{B}_0 \\
a_1^N \equiv a_1(k) \; &= (1/k)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_1^N)^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_1^N) + (\boldsymbol{b}_1^N - \boldsymbol{b}_0)^\top \boldsymbol{B}_0 (\boldsymbol{b}_1^N - \boldsymbol{b}_0) + a_0 \\
d_1^N \equiv d_1(k) \; &= n + d_0.
\end{aligned}$$

Thus, the marginal posterior for $\boldsymbol{\beta}$ is given by

$$p^N(\boldsymbol{\beta} \mid \boldsymbol{y}) = St_p(\boldsymbol{\beta} \mid \boldsymbol{b}_1^N, [\boldsymbol{T}_1^N]^{-1}, d_1^N), \tag{9}$$

where $\boldsymbol{T}_1^N = \left(\frac{d_1^N}{a_1^N}\right) \boldsymbol{B}_1^N$. It is then clear that $p^N(\boldsymbol{\beta}, \phi \mid \boldsymbol{y})$ depends upon k.

Reference prior and posterior The reference prior for $(\boldsymbol{\beta}, \phi)$ under the general normal model (7) is, again,

$$\pi(\boldsymbol{\beta}, \phi) \propto \phi^{-1},$$

regardless of the value of k. As in the standard case, this prior can be formally seen as the limit of the näive conjugate prior as $a_0 \to 0$, $d_0 \to -p$ and $\boldsymbol{B}_0 \to \boldsymbol{O}$.

Similarly, the reference posterior distribution is a limiting case of the corresponding conjugate posterior distribution and is characterized by

$$\begin{aligned}
\boldsymbol{b}_1^N &= \hat{\boldsymbol{\beta}} \\
\boldsymbol{B}_1^N &= (1/k)\boldsymbol{X}^\top \boldsymbol{X} \\
a_1^N &= (n-p)\,\hat{\sigma}_k^2 = (n-p)\,\hat{\sigma}^2/k \\
d_1^N &= n-p.
\end{aligned}$$

7.1.2 Compatible priors

As discussed at the end of in Section 6, the prior distribution should be elicited on a set of parameters having the same interpretation across models. There, as a general solution, we suggested working with quantiles. Alternatively, when the first two moments exist, one can work in terms of the mean and the variance.

Here it is possible to use priors on $(\boldsymbol{\beta}, \phi)$ which are compatible in the sense that they induce the same prior distribution on $\mathrm{E}(\boldsymbol{Y})$ and $\mathrm{Var}(\boldsymbol{Y})$. This is easily achieved in our case. Let $\tau = \phi/k$, so that $\mathrm{Var}(\boldsymbol{Y} \,|\, \boldsymbol{\beta}, \phi) = \tau^{-1}\boldsymbol{I}_n$. Then the prior

$$p_1^C(\boldsymbol{\beta}, \tau) = \mathrm{N}_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_0, \tau^{-1}\boldsymbol{B}_0^{-1})\ \mathrm{Ga}(\phi \,|\, d_0/2, a_0/2),$$

has the desired property. Written in terms of $(\boldsymbol{\beta}, \phi)$, this prior takes the form

$$p^C(\boldsymbol{\beta}, \phi) = \mathrm{N}_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_0^C, \phi^{-1}[\boldsymbol{B}_0^C]^{-1})\ \mathrm{Ga}(\phi \,|\, d_0^C/2, a_0^C/2),$$

with

$$\begin{aligned}
\boldsymbol{b}_0^C &= \boldsymbol{b}_0 \\
\boldsymbol{B}_0^C &= (1/k)\boldsymbol{B}_0 \\
a_0^c &= a_0/k \\
d_0^C &= d_0.
\end{aligned}$$

(Note that this prior coincides with the prior used in Sections 3.3 and 7.1.1 only when $k = 1$.) In other words,

$$p^C(\boldsymbol{\beta}, \phi) = p_k^C(\boldsymbol{\beta}, \phi) \equiv \mathrm{N}_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_0, (\phi/k)^{-1}\boldsymbol{B}_0^{-1})\ \mathrm{Ga}(\phi \,|\, d_0/2, a_0/(2k)).$$

The posterior distribution is given by

$$p^C(\boldsymbol{\beta}, \phi \,|\, \boldsymbol{y}) = \mathrm{N}_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_1^C, \phi^{-1}[\boldsymbol{B}_1^C]^{-1})\ \mathrm{Ga}(\phi \,|\, d_1^C/2, a_1^C/2),$$

where

$$\begin{aligned}
\boldsymbol{b}_1^C &= [(1/k)\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{B}_0^C]^{-1}[(1/k)\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{B}_0^C\boldsymbol{b}_0^C] \\
&= [(1/k)\boldsymbol{X}^\top\boldsymbol{X} + (1/k)\boldsymbol{B}_0]^{-1}[(1/k)\boldsymbol{X}^\top\boldsymbol{y} + (1/k)\boldsymbol{B}_0\boldsymbol{b}_0] \\
&= (\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{B}_0)^{-1}(\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{B}_0\boldsymbol{b}_0) \\
&= \boldsymbol{b}_1 \\
\boldsymbol{B}_1^C &= (1/k)\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{B}_0^C \\
&= (1/k)\boldsymbol{X}^\top\boldsymbol{X} + (1/k)\boldsymbol{B}_0 \\
&= (1/k)(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{B}_0) \\
&= (1/k)\boldsymbol{B}_1 \\
a_1^C &= (1/k)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_1^C)^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_1^C) + (\boldsymbol{b}_1^C - \boldsymbol{b}_0^C)^\top\boldsymbol{B}_0^C(\boldsymbol{b}_1^C - \boldsymbol{b}_0^C) + a_0^C \\
&= (1/k)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_1)^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_1) + (1/k)(\boldsymbol{b}_1 - \boldsymbol{b}_0)^\top\boldsymbol{B}_0(\boldsymbol{b}_1 - \boldsymbol{b}_0) + a_0/k \\
&= a_1/k \\
d_1^C &= n + d_0^C = d_1.
\end{aligned}$$

Note that this posterior can also be written as

$$p^C(\boldsymbol{\beta}, \phi \,|\, \boldsymbol{y}) = p_k^C(\boldsymbol{\beta}, \phi \,|\, \boldsymbol{y}) \equiv \mathrm{N}_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_1, (\phi/k)^{-1}\boldsymbol{B}_1^{-1})\,\mathrm{Ga}(\phi \,|\, d_1/2, a_1/(2k)).$$

Now, the marginal posterior for $\boldsymbol{\beta}$ is given by

$$p^C(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = St_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_1^C, [\boldsymbol{T}_1^C]^{-1}, d_1^C), \tag{10}$$

where $\boldsymbol{T}_1^C = \left(\frac{d_1^C}{a_1^C}\right)\boldsymbol{B}_1^C = \boldsymbol{T}_1$. In fact,

$$p^C(\boldsymbol{\beta} \,|\, \boldsymbol{y}) = St_p(\boldsymbol{\beta} \,|\, \boldsymbol{b}_1, \boldsymbol{T}_1^{-1}, d_1),$$

which does not depend on the value of k.

Reference prior and posterior Recall that the reference prior for $(\boldsymbol{\beta}, \phi)$ under the general normal model (7), is

$$\pi(\boldsymbol{\beta}, \phi) \propto \phi^{-1},$$

and note that it can also be formally regarded as the limit of the compatible conjugate prior when $a_0^C \to 0$, $d_0^C \to -p$ and $\boldsymbol{B}_0^C \to \boldsymbol{O}$, which of course is equivalent to $a_0 \to 0$, $d_0 \to -p$ and $\boldsymbol{B}_0 \to \boldsymbol{O}$.

Thus, the reference posterior distribution is characterized by

$$\begin{aligned}
\boldsymbol{b}_1^C &= \hat{\boldsymbol{\beta}} \\
\boldsymbol{B}_1^C &= (1/k)\boldsymbol{X}^\top\boldsymbol{X} \\
a_1^C &= (n-p)\,\hat{\sigma}_k^2 = (n-p)\,\hat{\sigma}^2/k \\
d_1^C &= n - p,
\end{aligned}$$

or, equivalently, by

$$\begin{aligned}\boldsymbol{b}_1 &= \hat{\boldsymbol{\beta}}\\ \boldsymbol{B}_1 &= \boldsymbol{X}^\top \boldsymbol{X}\\ a_1 &= (n-p)\,\hat{\sigma}^2\\ d_1 &= n-p.\end{aligned}$$

7.2 Inference on β

7.2.1 Näive prior

As discussed in Section 7.1.1, the marginal posterior of $\boldsymbol{\beta}$ under the general normal model (7), depends on the value of k and hence on the specific elliptical model under consideration. Indeed, Equation (9) can be very sensitive to the choice of k if the prior, Equation (8) is proper, even if such prior is rather vague.

To see this, suppose that $p^N(\boldsymbol{\beta}, \phi)$ is proper, so that $a_0 > 0$, $d_0 > 0$, and $\boldsymbol{B}_0$ is positive-definite. Then, as $k \to \infty$,

$$\begin{aligned}\boldsymbol{b}_1^N &\to \boldsymbol{b}_0\\ \boldsymbol{B}_1^N &\to \boldsymbol{B}_0\\ a_1^N &\to a_0\\ d_1^N &\to n + d_0,\end{aligned}$$

which implies

$$p^N(\boldsymbol{\beta}\,|\,\boldsymbol{y}) \to St_p\left(\boldsymbol{\beta}\,|\,\boldsymbol{b}_0, \left[\frac{a_0}{n+d_0}\right]\boldsymbol{B}_0^{-1}, n+d_0\right),$$

whereas, as $k \to 0$,

$$\begin{aligned}\boldsymbol{b}_1^N &\to \hat{\boldsymbol{\beta}}\\ k\boldsymbol{B}_1^N &\to \boldsymbol{X}^\top \boldsymbol{X}\\ ka_1^N &\to (n-p)\,\hat{\sigma}^2\\ d_1^N &\to n + d_0,\end{aligned}$$

formally implying

$$p^N(\boldsymbol{\beta}\,|\,\boldsymbol{y}) \to St_p\left(\boldsymbol{\beta}\,|\,\hat{\boldsymbol{\beta}}, \left[\frac{(n-p)\hat{\sigma}^2}{n+d_0}\right](\boldsymbol{X}^\top\boldsymbol{X})^{-1}, n+d_0\right).$$

It should be noted, however, that $k \to 0$ implies $\boldsymbol{Y} \to \boldsymbol{X}\boldsymbol{\beta}$, which in turn implies $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$ and $\hat{\sigma}^2 \to 0$. Hence, regardless of the values of the (fixed) sample size n, $p^N(\boldsymbol{\beta}\,|\,\boldsymbol{y})$ converges to a degenerate distribution at the true value of $\boldsymbol{\beta}$.

Summing up, even when the proper prior $p^N(\boldsymbol{\beta}, \phi)$ is arbitrarily close to reference prior, we can choose values of k leading to completely different inferences for $\boldsymbol{\beta}$. On this basis, it may be argued that the robustness (rather, invariance) obtained under the reference prior is just a singularity rather than a general pattern.

7.2.2 Compatible priors

In contrast with the results obtained for the näive prior, the marginal posterior for $\boldsymbol{\beta}$ under the general normal model (7), is the same for all values of k, and equals

$$p^C(\boldsymbol{\beta}\,|\,\boldsymbol{y}) = St_p(\boldsymbol{\beta}\,|\,\boldsymbol{b}_1, \boldsymbol{T}_1^{-1}, d_1),$$

the marginal posterior under the standard model (see Section 3.3).

Therefore, Equation (10) does not depend on the specific elliptical model under consideration, even if the prior $p^C(\boldsymbol{\beta}, \phi)$ is proper and quite informative. Of course, in the limiting case where $p^C(\boldsymbol{\beta}, \phi)$ is the reference prior, Equation (10), will still not depend on the choice of k. Thus, use of compatible priors (and not only the reference prior) will yield posterior inferences on $\boldsymbol{\beta}$ that are insensitive to the choice of k.

Concerning the effect of the prior on the marginal posterior of $\boldsymbol{\beta}$ for fixed k, it can be seen that the weight of the prior relative to the data is the same as in the standard linear model. This, again, is in contrast with the results obtained under the näive prior.

7.3 Prediction and model comparison

In this section we will only consider proper priors since the following predictive densities and the corresponding Bayes factors are not defined when reference priors are used.

7.3.1 Predictive densities

Predictive density under the standard model:

$$p(\boldsymbol{y}) = St_n\left(\boldsymbol{\beta}\,|\,\boldsymbol{X}\boldsymbol{b}_0, \left(\frac{a_0}{d_0}\right)(\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top), d_0\right).$$

Predictive density under the general normal model (näive prior):

$$p^N(\boldsymbol{y}) = St_n\left(\boldsymbol{\beta}\,|\,\boldsymbol{X}\boldsymbol{b}_0, \left(\frac{a_0}{d_0}\right)(k\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top), d_0\right).$$

Predictive density under the general normal model (compatible priors):

$$\begin{aligned} p^C(\boldsymbol{y}) &= St_n\left(\boldsymbol{\beta}\,|\,\boldsymbol{X}\boldsymbol{b}_0^N, \left(\frac{a_0^N}{d_0^N}\right)(k\boldsymbol{I}_n + \boldsymbol{X}[\boldsymbol{B}_0^N]^{-1}\boldsymbol{X}^\top), d_0^N\right) \\ &= St_n\left(\boldsymbol{\beta}\,|\,\boldsymbol{X}\boldsymbol{b}_0, \left(\frac{a_0}{d_0}\right)(\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top), d_0\right) \\ &= p(\boldsymbol{y}). \end{aligned}$$

7.3.2 Bayes factors

The Bayes factor between the standard model and the general normal model with a näive prior and the same design matrix is given by

$$\begin{aligned}
\mathrm{BF}_{12}^{N} = \mathrm{BF}_{12}^{N}(k) &= \frac{p(\boldsymbol{y})}{p^{N}(\boldsymbol{y})} \\
&= \frac{|k\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top|^{1/2}}{|\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top|^{1/2}} \\
&\times \frac{\left\{(1+\frac{1}{a_0}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_0)^\top(k\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_0)\right\}^{(d_0+p)/2}}{\left\{(1+\frac{1}{a_0}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_0)^\top(\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_0)\right\}^{(d_0+p)/2}}
\end{aligned}$$

Since the prior is proper, we must have $a_0 > 0$, $d_0 > 0$, and $\boldsymbol{B}_0$ positive-definite. Now, as $k \to \infty$,

$$\mathrm{BF}_{12}^{N} = \mathrm{BF}_{12}^{N}(k) \to \infty,$$

while, letting $k \to 0$, we have

$$\mathrm{BF}_{12}^{N} = \mathrm{BF}_{12}^{N}(k) \to \frac{|\boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top|^{1/2}}{|\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top|^{1/2}} \times$$

$$\frac{\left\{(1+\frac{1}{a_0}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_0)^\top(\boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_0)\right\}^{(d_0+p)/2}}{\left\{(1+\frac{1}{a_0}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_0)^\top(\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{B}_0^{-1}\boldsymbol{X}^\top)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_0)\right\}^{(d_0+p)/2}} \;<\; 1.$$

Thus, under a proper näive prior it is possible to choose different values of k leading to quite different Bayes factors, and so the analysis is not robust against the choice of k.

On the other hand, to study the sensitivity of BF_{12}^{N} to the choice of (näive) prior for fixed k, for simplicity we consider the case where $B_0 = v_0\boldsymbol{I}_p$ with $0 < v_0 < \infty$. Then, it can be shown that

$$\lim_{v_0 \to \infty} \mathrm{BF}_{12}^{N}(k) > k^{n/2} \quad (\text{if } k > 1)$$

and

$$\lim_{v_0 \to \infty} \mathrm{BF}_{12}^{N}(k) < k^{n/2} \quad (\text{if } k < 1).$$

(Note that $\mathrm{BF}_{12}^{N}(1) = 1$ for all v_0.) Also, note that

$$\lim_{v_0 \to 0} \mathrm{BF}_{12}^{N}(k) = 1,$$

for each fixed value of k.

We now turn our attention to the Bayes factor between the standard model and the general normal model with a compatible prior and the same design matrix. It is given by

$$\mathrm{BF}_{12}^{C} = \frac{p(\boldsymbol{y})}{p^{C}(\boldsymbol{y})} = 1, \forall \; k.$$

Therefore, when compatible priors are used the general normal model (for all values of k) is equivalent to the standard model. This makes sense in our case since the general normal model is just a rescaled version of the standard normal model and only affects the variance of the observations, while the effect of the compatible prior is precisely to compensate for such a change in the scale by assigning the same prior to the variance of the observations regardless of the sampling model. It should be noted however, that when comparing two general elliptical linear models, BF_{12}^{C} should not be insensitive as in this case but should be able to adequately discriminate between them if the corresponding generators belong to different families with different tail behavior or kurtosis, say.

Concerning the effect of the prior on the Bayes factor for fixed k, it can be seen that the Bayes factor is insensitive to the choice of prior in this case. This, once more, is in contrast with the results obtained under the näive prior. However, as discussed above, when comparing two general elliptical linear models, BF_{12}^{C} should not be insensitive as in this case but should be robust against changes in the prior.

Here we are not concerned with the problem of model comparison under reference (improper) priors due to the fact that Bayes factors are not well defined and no widely accepted solution exists in that case. However, the analysis above suggests that suitably defined limits of Bayes factors under compatible priors may be used to compare general elliptical model in the noninformative case.

8. Concluding Remarks

In this paper, we have argued that the invariance property exhibited by regression and location-scale models, and widely discussed in the literature, is a rather extreme instance of robustness, in fact verging on complete insensitivity (Figure 1).

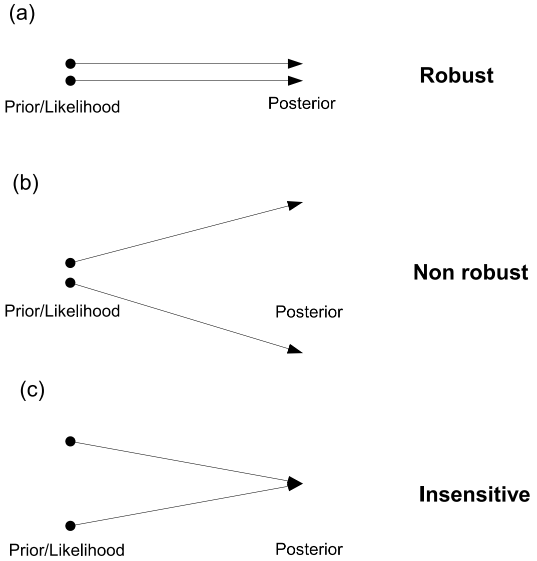

Figure 1. Schematic representation of Bayesian inference as regards robustness.

From this perspective, it is not at all clear that such a property is a desirable one. One important aspect of this property is the need for the use of the same diffuse (improper) prior distribution on the location and scale parameters across sampling models, even though the interpretation of such parameters vary.

We have considered a case study focusing specifically on the robustness of inferences concerning regression coefficients $\boldsymbol{\beta}$ of the normal linear model with respect to changes in two different inputs: (i) the generator $h \in \mathcal{H}$ for a very simple class of elliptical sampling models $\mathcal{H}$; and (ii) the prior distribution on the regression and scale parameters.

Specifically, we have studied how small changes in h or in the prior affect the marginal posterior distribution of the regression coefficients $\boldsymbol{\beta}$. Roughly speaking, if such small changes lead to small changes in the posterior then the inferences about $\boldsymbol{\beta}$ can actually said to be robust (Figure 1(a)). Conversely, if small changes in h or in the prior produced significant changes in the posterior, then the inferences about $\boldsymbol{\beta}$ would not be robust (Figure 1(b)).

We can identify a third possibility, namely the situation when changes (large or small) in h or in the prior do not change the posterior distribution of $\boldsymbol{\beta}$ at all (Figure 1(c)). We find this lack of sensitivity quite intriguing and feel that it should be regarded as essentially different from the robustness property described above and represented by (Figure 1(a)). Similar remarks can be made about the robustness of the predictive (*i.e.* marginal) distributions of the data and of the Bayes factors derived from them.

While we acknowledge we have focused on a very simple model, it does allow us to compute all the required densities (posterior and predictive) in closed form while providing some insight concerning the subtleties of robustness analyses for more general location-scale models. For one thing, use of näive priors should be avoided and use of compatible priors is recommended. We also feel that a proper robustness study should concentrate on proper (compatible) priors and avoid the use of diffuse (improper) prior.

Indeed, under proper compatible priors, posterior inferences on $\boldsymbol{\beta}$ are insensitive to the choice of h within the class of rescaled normal generators $\mathcal{H}$. This result is closely related to that obtained by Arellano-Valle et al. (2003) within the general class of elliptical models when proper conjugate priors are used. Similarly, under proper compatible priors, all of the general normal models derived from the class $\mathcal{H}$ are equivalent according to the Bayes factor. However, the Bayes factor can be expected to distinguish between two different general elliptical linear models with the same design matrix.

As a final remark, it must be pointed out that we have not discussed a number of contributions where asymptotic calculations are part of the central arguments (see, e.g., Gustafson, 2001 and Copas and Eguchi, 2010). This is because, despite the theoretical value of those results, for most practical robustness studies the sample size will be moderate at best.

Acknowledgements

Robustness is but one of the many subjects where the legacy of Pilar will persist through the papers she wrote along with a number of colleagues, students and friends. We are certain that she would be very enthusiastic about reading comments on her work, and we are also convinced that she would react with a vigorous reply. Sadly, she is no longer with us to enrich our thoughts with her views. We present this paper as a tribute to her memory.

This work was partially supported by Sistema Nacional de Investigadores, México. The first author wishes to acknowledge support from Asociación Mexicana de la Cultura, A.C.

Appendix: Elliptical Distributions

A real, absolutely continuous random variable Z is said to have a spherical distribution if it has a density function of the form

$$f(z) = h(z^2) \quad (z \in \mathrm{I\!R}),$$

where the function $h(\cdot)$ is called the generator and is such that

$$\int_0^\infty u^{-1/2}\, h(u)\, \mathrm{d}u = 1.$$

For example, if $h(u) = (2\pi)^{-1/2} \exp(-u/2)$ then $Z \sim \mathrm{N}(0,1)$ and we obtain the standard normal distribution. Other choices of $h(\cdot)$ give rise to other spherical distribution such as the Student-t. Multivariate spherical distributions for an n-dimensional vector $\boldsymbol{Z}$ can be similarly defined by

$$f(\boldsymbol{z}) = h^n(||\boldsymbol{z}||^2) \quad (z \in \mathrm{I\!R}^n),$$

where $||z||^2 = \boldsymbol{z}^\top \boldsymbol{z}$ and the generator $h^n(\cdot)$ is such that

$$\int_0^\infty \frac{\pi^{n/2}}{\Gamma(n/2)} u^{(n/2)-1}\, h^n(u)\, \mathrm{d}u = 1.$$

More generally, a random vector $\boldsymbol{Y}$ is said to have an elliptical distribution with location vector $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$ is its density function is of the form

$$f(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-1/2} h^n((\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})).$$

This is commonly written as $\boldsymbol{Y} \sim \mathrm{EL}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; h)$.

For example, the multivariate normal model corresponds to

$$h^n(u) = (2\pi)^{-n/2} \exp(-u/2).$$

In this paper we will focus on a specific subclass given by generators of the form

$$h^n(u; k) = (2\pi k)^{-n/2} \exp(-u/(2k)).$$

We call this the general normal class and denote it by

$$\mathcal{H} = \{h^n(\cdot; k) : 0 < k < \infty\}.$$

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W., 1972. Robust Estimates of Location: Survey and Advances. Princeton University Press, Princeton.

Arellano-Valle, R.B., Galea-Rojas, M., Iglesias, P., 2000. Bayesian sensitivity analysis in elliptical linear regression models. Journal of Statistical Planning and Inference, 86, 175-199.

Arellano-Valle, R.B., Iglesias, P.L., Vidal, I., 2003. Bayesian inference for elliptical models: conjugate analysis and model comparison (with discussion). In Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M., (eds.). Bayesian Statistics 7. University Press, Oxford, pp. 3-24.

Arellano-Valle, R.B., del Pino, G., Iglesias, P., 2006. Bayesian inference in spherical linear models: robustness and conjugate analysis. Journal of Multivariate Analysis, 97, 179-197.

Azzalini, A., 1985. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics, 12, 171-178.

Berger, J.O., 1990. Robust Bayesian analysis: sensitivity to the prior. Journal of Statistical Planning and Inference, 25, 303-328.

Berger, J.O., 1994. An overview of robust Bayesian analysis (with discussion). Test, 3, 5-124.

Berger, J.O., Ríos-Insua, D., Ruggeri, F., 2000. Bayesian Robustness. In Ríos-Insua, D., Ruggeri, F., (eds.). Robust Bayesian Analysis. Springer-Verlag, New York.

Box G.E.P., 1953. Non-normality and tests on variances. Biometrika, 40, 318-335.

Box, G.E.P, Tiao, G.C., 1962. A further look at robustness via Bayes's Theorem. Biometrika, 49, 419-432.

Copas, J., Eguchi, S., 2010. Likelihood for statistically equivalent models. Journal of The Royal Statistical Society Series B—Statistical Methodology, 72, 193-217.

de Finetti, B., 1974. Bayesianism: its unifying role for both the foundations and applications of statistics. International Statistical Review, 42, 117-130.

Dickey, J.D., Chen, C.-H., 1985. Direct Subjective-Probability Modelling Using Ellipsoidal Distributions. In Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M., (eds.). Bayesian Statistics 2. Elsevier, Amsterdam, pp. 157-182.

Fernández, C., Osiewalski, J., Steel, M.F.J., 1994. The continuous multivariate location-scale model revisited: a tale of robustness. Biometrika, 81, 558-594.

Fernández, C., Osiewalski, J., Steel, M.F.J., 1995. Modeling and inference with v-spherical distributions. Journal of the American Statistical Association, 90, 1331-1340.

Fernández, C., Osiewalski, J., Steel, M.F.J., 1997. Classical and Bayesian inference robustness in multivariate regression models. Journal of the American Statistical Association, 92, 1434-1444.

Gustafson, P., 2001. On measuring sensitivity to parametric model misspecification. Journal of The Royal Statistical Society Series B—Statistical Methodology, 63, 81-94.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

Hoaglin, D.C., Mosteller, F., Tukey, J.W., (eds.) 1983. Understanding Robust and Exploratory Data Analysis. Wiley, New York.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Huber, P.J., 2002. John W. Tukey's contributions to robust statistics. The Annals of Statistics, 30, 1640-1648.

Huber, P.J., 2009. Robust Statistics. (2nd. ed.). Wiley, New York.

Morgenthaler, S., 2007. A survey of robust statistics. Statistical Methods and Applications, 15, 271-293.

Morgenthaler, S., Tukey, J.W., (eds.), 1991. Configural Polysampling: a Route to Practical Robustness. Wiley, New York.

Osiewalski, J., 1991. A note on Bayesian inference in a regression model with elliptical errors. Journal of Econometrics, 48, 183-193.

Osiewalski, J., Steel, M.F.J., 1993a. Robust Bayesian inference in elliptical regression models. Journal of Econometrics, 57, 345-363.

Osiewalski, J., Steel, M.F.J., 1993b. Bayesian marginal equivalence of elliptical regression models. Journal of Econometrics, 59, 391-403.

Osiewalski, J., Steel, M.F.J., 1993c. Robust Bayesian inference in l_q-spherical models. Biometrika, 80, 456-460.

Ríos-Insua, D., Ruggeri, F., (eds.) , 2000. Robust Bayesian Analysis. Springer-Verlag, New York.

Tukey, J.W. 1960. A Survey of Sampling from Contaminated Distributions. In Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., Mann, H.B., (eds.). Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. University Press, Stanford, pp. 448-485.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading.

Vidal, I., Iglesias, P., Branco, M.D., Arellano-Valle, R.B., 2006. Bayesian sensitivity analysis and model comparison for skew elliptical models. Journal of Statistical Planning and Inference, 136, 3435-3457.

Zellner, A., 1976. Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. Journal of the American Statistical Association, 71, 400-405.