CLASSICAL AND BAYESIAN INFERENCE ROBUSTNESS IN MULTIVARIATE REGRESSION MODELS

Carmen Fernández, Jacek Osiewalski and Mark F.J. Steel

ABSTRACT

We consider a one-to-one correspondence between points $z \in \Re^n - \{0\}$ and pairs (y, r), where $r > 0$ and y lies in some space $\mathcal{Y}$, through $z = ry$. As an immediate consequence, we can represent random variables Z that take values in $\Re^n - \{0\}$ as $Z = RY$, where R is a positive random variable and Y takes values in $\mathcal{Y}$.

By fixing the distribution of either R or Y, while imposing independence between them, we generate classes of distributions on $\Re^n$. Many families of multivariate distributions, like *e.g.* spherical, elliptical, l_q-spherical, v-spherical and anisotropic, can be interpreted in this unifying framework.

Some classical inference procedures can be shown to be completely robust in these classes of multivariate distributions. These findings are used in the practically relevant context of regression models. Finally, we present a robust Bayesian analysis and indicate the links between classical and Bayesian results. In particular, for the regression model with i.i.d. errors up to a scale, a formal characterization is provided for both classical and Bayesian robustness results concerning inference on the regression parameters.

KEY WORDS: Distribution-invariance; Independent sampling; Likelihood ratio principle; Pivotal quantity; Posterior inference; Scale invariance.

Carmen Fernández was Research Associate at the Institut de Statistique, Université Catholique de Louvain when this research started, and is now a Research Fellow at the CentER for Economic Research and Assistant Professor at the Department of Econometrics, Tilburg University, 5000 LE Tilburg, The Netherlands. Jacek Osiewalski is Associate Professor at the Department of Econometrics, Academy of Economics, 31-510 Kraków, Poland and was visiting CORE, Université Catholique de Louvain, when the work on this paper began. Mark F.J. Steel is Senior Research Fellow at the CentER for Economic Research and Associate Professor at the Department of Econometrics, Tilburg University. The first and third authors acknowledge support from the Netherlands Organization for Scientific Research (NWO), and the second author was partly supported by the Belgian Programme on Interuniversity Poles of Attraction. Helpful comments from Julián de la Horra, Michel Mouchart, Jean-Marie Rolin, an associate editor and a referee are gratefully acknowledged.

1. INTRODUCTION

This paper investigates perfectly robust inference from vector observations. We are particularly interested in examining whether any parallels can be found between classical and Bayesian robust inference results.

We shall call an inference procedure robust (or distribution-invariant) if it is not affected by changes in the sampling distribution over a particular class. Thus, we are not considering robustness with respect to the data (extreme observations), or with respect to the prior specification in a Bayesian framework, but we focus on exact robustness with respect to the specification of the sampling model. In particular, we analyze what Box and Tiao (1973, p.152) call "criterion robustness" in a classical framework, whereas Bayesian results relate to "model robustness" [see Berger (1985, p.248)].

In order to provide us with a natural way to define classes of sampling distributions over which we investigate robustness, Section 2 introduces a representation of points in $\Re^n - \{0\}$ in terms of pairs (y, r), where $r > 0$ and y lies in some $(n-1)$-dimensional space $\mathcal{Y}$. This representation can be thought of as a generalization of the usual polar coordinates, where r is the Euclidean norm and $\mathcal{Y}$ is the unit sphere S^{n-1}. A number of examples show that especially the classes generated by fixing the distribution of Y while imposing independence between R and Y are practically useful. However, through fixing the distribution of R or relaxing the independence constraint other types of classes are naturally generated.

Inference is often conducted on the basis of several vector observations, usually the result of independent sampling. We explicitly discuss inference in such a repeated sampling framework, where we focus on independent and identically distributed (i.i.d.) error vectors, up to a scale. Section 3 presents results characterizing distribution-invariant functions of matrix random variables, and applies these findings to robust classical inference in the context of regression models with scale. A parallel robust Bayesian analysis of these models is described in Section 4. For both paradigms, the robustness results essentially hinge upon the presence of a scale parameter for each vector observation.

The final Section groups some conclusions and puts the similarities between both classical and Bayesian approaches in perspective. All proofs are referred to the Appendix, without explicit mention in the main text.

2. CLASSES OF MULTIVARIATE DISTRIBUTIONS

2.1. A representation of $\Re^n$

In this Subsection, we shall introduce a general representation of points in $\Re^n$. This will prove to be a useful tool for generating and analyzing multivariate distributions and naturally induces useful classes of distributions in which robustness results can be derived.

We shall establish a one-to-one correspondence between points $z \in \Re^n - \{0\}$ and pairs $(y, r) \in \mathcal{Y} \times \Re_+$, such that $z = ry$. This implies that $\mathcal{Y}$ is an $(n-1)$-dimensional manifold that can be expressed as $\mathcal{Y} = \{k(w) : w \in \mathcal{W}\}$, where $\mathcal{W} \subset \Re^{n-1}$ is the set of angular polar coordinates and $k(\cdot)$ is a function such that the angular polar coordinates of $k(w)$ are given by w. Thus, we can also uniquely identify $z \in \Re^n - \{0\}$ with a pair $(w, r) \in \mathcal{W} \times \Re_+$, such that $z = rk(w)$.

In this representation, different choices of $\mathcal{Y}$ [alternatively, different choices of $k(\cdot)$] lead to different interpretations of r and y. For instance, if $\mathcal{Y}$ is the unit sphere S^{n-1}, we obtain $y = z/\|z\|_2 \in S^{n-1}$ and $r = \|z\|_2$, the Euclidean radius. This leads to the usual polar representation. More generally, we could take *e.g.* $\mathcal{Y} = \{x \in \Re^n : \|x\|_q \equiv (\sum_{i=1}^n |x_i|^q)^{1/q} = 1\}$, the unit l_q-sphere, for some $q \in [1, \infty)$, in which case $r = \|z\|_q$ describes the l_q-radius, whereas $y = z/\|z\|_q$ is the corresponding point on the unit l_q-sphere.

This immediately provides us with a representation for random variables Z that take values in $\Re^n - \{0\}$, as

$$Z = RY, \tag{2.1}$$

where R is a positive random variable and Y takes values in $\mathcal{Y}$. Furthermore, we can identify Y with a random variable W taking values in $\mathcal{W} \subset \Re^{n-1}$ through $Y = k(W)$.

In this way, for a given choice of $\mathcal{Y}$, we have established a one-to-one correspondence between any random variable Z that takes values in $\Re^n - \{0\}$ and a pair (Y, R) or, equivalently, (W, R). The correspondence between the distributions of Z and (Y, R) naturally leads to the definition of classes of distributions in $\Re^n - \{0\}$ characterized by fixing the marginal distribution of either R or Y. Many well-known families of multivariate distributions can be generated in this way, usually assuming independence between R and Y, as is illustrated in the remainder of this Section.

2.2. Fixing the distribution of Y

Given a choice of $\mathcal{Y}$ and any probability distribution P on $\mathcal{Y}$, we define the class $\mathcal{S}$ as the following set of random variables on $\Re^n - \{0\}$

$$\mathcal{S} = \{Z : Z = RY \text{ through (2.1), with } R \text{ and } Y \text{ independent and } Y \text{ distributed as } P\}. \tag{2.2}$$

We now consider the class of distributions corresponding to the random variables in $\mathcal{S}$. By varying the choice of $\mathcal{Y}$ and P, we generate many useful classes of multivariate distributions. As an important example, the well-known class of n-variate **spherical distributions** [as in Kelker (1970)] corresponds to $\mathcal{S}$ with $\mathcal{Y}$ the unit sphere S^{n-1} and Y uniformly distributed over S^{n-1} [see Fang, Kotz and Ng (1990)]. We now show that the classes of **υ-spherical distributions**, introduced in Fernández, Osiewalski and Steel (1995), can also be interpreted in the framework of (2.2). Fernández *et al.* (1995) define a continuous υ-spherical distribution through the following probability density function (p.d.f.) of $Z = (Z_1, \ldots, Z_n)'$ in $\Re^n$:

$$p(z) = g\{\upsilon(z)\}, \tag{2.3}$$

where $\upsilon(\cdot)$ is a scalar function, strictly positive for $z \in \Re^n - \{0\}$ and such that $\upsilon(\alpha z) = \alpha\upsilon(z)$ for all $\alpha > 0$ and $z \in \Re^n$, and $g(\cdot)$ is a nonnegative labelling function. Furthermore, $\upsilon(\cdot)$ and $g(\cdot)$ are such that (2.3) is proper. For every admissible choice of $\upsilon(\cdot)$, the corresponding class of continuous υ-spherical distributions is obtained by allowing $g(\cdot)$ to be free.

The name "υ-spherical" is motivated by the fact that all members of the same class have common isodensity sets of the form $\{z \in \Re^n : \upsilon(z) = \eta\}$, where η is a positive constant. We shall call the latter set the "υ-sphere" of "υ-radius" η. Through judicious choices of $\upsilon(\cdot)$ we can accommodate a wide range of possible isodensity contours. Clearly, continuous spherical distributions are a special case of continuous υ-spherical distributions where $\upsilon(\cdot)$ is the Euclidean norm, whereas continuous **l_q-spherical distributions**, introduced in Osiewalski and Steel (1993), correspond to the l_q-norm (or a generalization with $0 < q < 1$) for $\upsilon(\cdot)$. Another interesting example are the **elliptical distributions**, which correspond to $\upsilon(z) = (z'V^{-1}z)^{1/2}$ for any positive definite symmetric matrix V. Fernández *et al.* (1995) provide several non-standard examples of υ-spherical classes.

We now show that any υ-spherical class is a special case of $\mathcal{S}$ in (2.2) where $\mathcal{Y}$ is the unit υ-sphere. Note that our assumptions on $\upsilon(\cdot)$ imply that the unit υ-sphere is a valid choice for $\mathcal{Y}$. Transforming from $z \in \Re^n - \{0\}$ to (w, r), where $w = h^{-1}(z/\|z\|_2) \in \mathcal{W} \subset \Re^{n-1}$ denotes the angular polar coordinates and $r = \upsilon(z) > 0$ is the υ-radius, we obtain from (2.3)

$$p(w,r) = s(w)[\upsilon\{h(w)\}]^{-n} r^{n-1} g(r), \tag{2.4}$$

where $s(w)$ is the function of w that appears in the Jacobian of the transformation from the Cartesian to the polar coordinates of z. From (2.4) we immediately deduce

$$p(w) \propto s(w)[\upsilon\{h(w)\}]^{-n} \quad \text{and} \quad p(r|w) = p(r) \propto r^{n-1} g(r), \tag{2.5}$$

which implies independence between the polar angles $W = h^{-1}(Z/\|Z\|_2)$ and the υ-radius $R = \upsilon(Z)$. In addition, only $\upsilon(\cdot)$ determines the distribution of W (or, equivalently, the distribution over the unit υ-sphere), whereas we can accommodate any proper density for R, since we have no restrictions on $g(\cdot)$ other than properness of $p(w,r)$ in (2.4). Thus, for each given $\upsilon(\cdot)$, the class of continuous υ-spherical distributions corresponds to a subset of a particular $\mathcal{S}$ in (2.2), where Y takes values in the unit υ-sphere with its distribution fixed by the choice of $\upsilon(\cdot)$ through $p(w)$ in (2.5), and the υ-radius R has any p.d.f. on $\Re_+$. If we extend this class of distributions by also allowing for any noncontinuous distribution for R, we cover the entire class $\mathcal{S}$.

2.3. Fixing the distribution of R

Given a particular choice of $\mathcal{Y}$, and a fixed probability distribution Q on $\Re_+$, we define the class of random variables on $\Re^n - \{0\}$

$$\mathcal{R} = \{Z : Z = RY \text{ through (2.1), with } R \text{ and } Y \text{ independent and } R \text{ distributed as } Q\}. \tag{2.6}$$

These classes of distributions have received much less attention in the literature than their conterparts based on $\mathcal{S}$ in (2.2). The only example that we are aware of is the class of **anisotropic distributions**, introduced in Nachtsheim and Johnson (1988), which exactly fits into the framework of (2.6), by choosing $\mathcal{Y} = S^{n-1}$ and a $\sqrt{\chi^2_n}$ distribution for the Euclidean radius $R = \|Z\|_2$. Note that Normality, which corresponds to uniformity of $Y = Z/\|Z\|_2$ on S^{n-1}, defines the intersection of the anisotropic and spherical classes. Like sphericity, the anisotropic family can be used to represent departures from Normality, although it constitutes a generalization in a complementary direction. Whereas the continuous spherical class allows for any labelling function while retaining the same isodensity sets, continuous anisotropic distributions do not preserve the shape of the isodensity sets, as the examples in Nachtsheim and Johnson (1988) illustrate.

3. ROBUSTNESS OF CLASSICAL INFERENCE

This Section combines results on distribution theory with inference in a sampling theory context. From a practical perspective, the most useful classes are generated as $\mathcal{S}$ in (2.2). Therefore, the discussion will be focused on these classes. Since inference is typically conducted on the basis of more than one vector observation, we shall consider matrices of observables, rather than single vectors. Subsection 3.1 derives results on distribution-invariant functions of random variables for a matricvariate generalization of $\mathcal{S}$ (without yet introducing parameters into the sampling model), whereas Subsection 3.2 applies this theory to robust inference on the location of a regression model in a context of independent sampling from $\mathcal{S}$. There is a substantial literature on distributional properties and robustness of classical inference under matrix random variables; however, it is typically only within the context of sphericity or ellipticity and outside the framework of independent sampling, which is our main focus here. Fang and Zhang (1990) provide a good overview of this literature.

3.1. Distribution invariance results in $\mathcal{S}$

We now focus on $(n \times p)$-dimensional matrix random variables $\mathbf{Z} = (Z_1, \dots, Z_p)$ where each column Z_i takes values in $\Re^n - \{0\}$ and is represented as in (2.1) through $Z_i = R_i Y_i$, with R_i a positive random variable and Y_i taking values in some $\mathcal{Y}_i$. Then we define

$$\mathcal{MS} = \{\mathbf{Z} = (Z_1, \dots, Z_p) : \mathbf{Z} = (R_1 Y_1, \dots, R_p Y_p) \text{ with } (R_1, \dots, R_p) \text{ independent of } (Y_1, \dots, Y_p) \text{ and the distribution of } (Y_1, \dots, Y_p) \text{ fixed}\}. \tag{3.1}$$

Thus, for given $\mathcal{Y}_1, \dots, \mathcal{Y}_p$ and a fixed distribution of $(Y_1, \dots, Y_p)$, the class $\mathcal{MS}$ is generated by considering all possible distributions for $(R_1, \dots, R_p)$, while imposing independence between $(R_1, \dots, R_p)$ and $(Y_1, \dots, Y_p)$. Clearly, in the case $p = 1$, the class $\mathcal{MS}$ reduces to our earlier $\mathcal{S}$ for vector distributions in (2.2). Also note that if $\mathcal{Y}_i = \mathcal{Y}$ for all $i = 1, \dots, p$, and the fixed joint distribution of $(Y_1, \dots, Y_p)$ leads to the same marginal distribution for each Y_i on $\mathcal{Y}$, then each of the columns of $\mathbf{Z}$ belongs to the same class $\mathcal{S}$.

A class in (3.1) that has appeared in the literature is generated by choosing $\mathcal{Y}_i = S^{n-1}$, $i = 1, \dots, p$, and fixing $Y_1, \dots, Y_p$ to have independent uniform distributions on S^{n-1}. This is the class of multivariate spherical distributions as studied in Fang and Zhang (1990) and corresponds to sphericity for $p = 1$. The class of vector-spherical distributions [Fang and Zhang (1990, Subsection 3.1.4)] is a subclass of the latter, where, in addition, we fix $(R_1^2/R^2, \dots, R_p^2/R^2)$ to be Dirichlet$(n/2, \dots, n/2)$ distributed, independently of $R^2 = \sum_{i=1}^p R_i^2$. Multivariate spherical distributions are, in turn, a subset of left-spherical distributions, first defined by Dawid (1977) and reviewed by Fang and Zhang (1990, Subsection 3.1.1). Transforming the random variables $\mathbf{Z}$ in all of these classes to $\mathbf{Z}^* = B\mathbf{Z}$, where B is some known $l \times n$ matrix of rank l, we naturally generate elliptical counterparts [see Fang and Zang (1990, Section 3.3)]. Representing $Z_i^* = R_i^* Y_i^*$, $i = 1, \dots, p$, where Y_i^* lies on the $(l-1)$-dimensional unit v-sphere with $v(z) = \{z'(BB')^{-1}z\}^{1/2}$ (which is an ellipsoid) and R_i^* is the corresponding v-radius, we can show that the elliptical counterpart of multivariate sphericity is again a class $\mathcal{MS}$ with $Y_1^*, \dots, Y_p^*$ i.i.d. with p.d.f. of the polar angles given through $p(w)$ in (2.5) for this choice of $v(\cdot)$. Thus, the class $\mathcal{MS}$ in (3.1) can be regarded as a generalization of the multivariate spherical class, or its elliptical counterpart, by means of the general representation introduced in (2.1). As we shall see in the sequel, $\mathcal{MS}$ will be useful for examining robustness of inference procedures under independent sampling from $\mathcal{S}$.

We now consider functions of the random variable $\mathbf{Z}$ that take values in $\Re^k$, *i.e.*, say, $t(\mathbf{z}) \in \Re^k$. In case $\mathbf{Z}$ describes a sampling process, such functions are usually called statistics. The next Theorem characterizes the functions that are distribution-invariant in $\mathcal{MS}$, in the sense that $t(\mathbf{Z})$ has the same distribution for all $\mathbf{Z} = (Z_1, \ldots, Z_p)$ in $\mathcal{MS}$.

Theorem 1. *Let $\mathcal{MS}$ be the class defined in* (3.1) *and $t(\cdot)$ any measurable function from $\Re^{n\times p}$ to $\Re^k$. Then*

$$t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p), \text{ for all } (Z_1, \ldots, Z_p) \in \mathcal{MS} \tag{3.2}$$

if and only if

$$t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p), \text{ for all } \alpha_1, \ldots, \alpha_p > 0 \text{ and for all } (Z_1, \ldots, Z_p) \in \mathcal{MS}. \tag{3.3}$$

While we present (3.2) in terms of a reference case $(Y_1, \ldots, Y_p)$, which is the element in $\mathcal{MS}$ corresponding to a Dirac distribution on $(1, \ldots, 1)$ for $(R_1, \ldots, R_p)$, we could alternatively state (3.2) as $t(\mathbf{Z}) \stackrel{d}{=} t(\mathbf{Z}^*)$, for all $\mathbf{Z}, \mathbf{Z}^* \in \mathcal{MS}$. Thus, Theorem 1 tells us that a function $t(\mathbf{Z})$ is distribution-invariant in $\mathcal{MS}$ if and only if its distribution is unaffected by changing the scales of each of the columns of $\mathbf{Z}$.

Let us now review some special cases of Theorem 1 that have appeared in the literature. Theorem 1 specializes to Theorem 5.1.1 (b) in Fang and Zhang (1990) for $\mathcal{MS}$ the multivariate spherical class, whereas it reduces to Theorem 7.3 of Fang *et al.* (1990) in the case of one vector observation ($p = 1$). Formally, their Theorem 7.3 is derived for random variables in a radially decomposable class [see Ng and Fraser (1994)], *i.e.* where Z has a stochastic representation $Z \stackrel{d}{=} RY$, with R a positive random variable independent of Y, and the latter has a fixed distribution, but without imposing a one-to-one correspondence between points $z \in \Re^n - \{0\}$ and pairs $(y, r) \in \mathcal{Y} \times \Re_+$ as introduced in Section 2. We could trivially extend Theorem 1 to this more general framework, as the proof does not rely on this one-to-one correspondence. Our reason for adopting it is to convey an interpretation to the random variables R and Y, which seems instrumental in the use of our framework for practical modelling purposes, as was illustrated through the classes discussed in Section 2.

As already mentioned, we shall typically be interested in matricvariate random variables in $\mathcal{MS}$ as a way of representing independent sampling from random vectors described through $\mathcal{S}$. To aid in the discussion of these issues, let us present the following proposition.

Proposition 1. *Consider any measurable function $t(\cdot)$ from $\Re^{n\times p}$ to $\Re^k$.*
(i) Let $\mathcal{MS}_{IS}$ be the subclass of $\mathcal{MS}$ corresponding to $R_1, \ldots, R_p$ all independent and such that $R_i \stackrel{d}{=} c_i R_1$ for some scalar $c_i > 0$, $i = 2, \ldots, p$. Then

$$t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p), \text{ for all } \mathbf{Z} \in \mathcal{MS}_{IS} \tag{3.4}$$

if and only if

$$t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p), \text{ for all } \alpha_1, \ldots, \alpha_p > 0 \text{ and for all } \mathbf{Z} \in \mathcal{MS}_{IS}. \quad (3.5)$$

(ii) Let $\mathcal{MS}_{IID}$ *be the subclass of* $\mathcal{MS}$ *corresponding to* $R_1, \ldots, R_p$ *i.i.d. Then*

$$t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p), \text{ for all } \alpha_1, \ldots, \alpha_p > 0 \text{ and for all } \mathbf{Z} \in \mathcal{MS}_{IID} \quad (3.6)$$

implies

$$t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p), \text{ for all } \mathbf{Z} \in \mathcal{MS}_{IID} \quad (3.7)$$

which, in turn, implies

$$t(\alpha Z_1, \ldots, \alpha Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p), \text{ for all } \alpha > 0 \text{ and for all } \mathbf{Z} \in \mathcal{MS}_{IID}. \quad (3.8)$$

(iii) Moreover, conditions (3.4), (3.5) *and* (3.6) *are all equivalent to* (3.2).

Proposition 1 (i) relates to the interesting case of sampling independently from the same distribution but with possibly different scales. If we choose $\mathcal{MS}$ such that $Y_1, \ldots, Y_p$ are i.i.d., then for each $\mathbf{Z} = (Z_1, \ldots, Z_p) \in \mathcal{MS}_{IS}$, $Z_1, \ldots, Z_p$ are random vectors from the corresponding class $\mathcal{S}$ and are i.i.d. up to a scale. Thus, the characterization of distribution-invariance with this type of sampling from $\mathcal{S}$ is given by (3.5).

If we go to another practically relevant case, namely i.i.d. sampling, we focus on Proposition 1 (ii). Choosing $\mathcal{MS}$ such that $Y_1, \ldots, Y_p$ are i.i.d., the columns $Z_1, \ldots, Z_p$ of $\mathbf{Z} \in \mathcal{MS}_{IID}$ are i.i.d. random vectors from the corresponding $\mathcal{S}$. Therefore, (3.6) provides a sufficient condition and (3.8) gives a necessary condition for $t(Z_1, \ldots, Z_p)$ to be distribution-invariant under i.i.d. sampling from $\mathcal{S}$. Since (3.6) and (3.8) are not equivalent, we do not have a characterization of all distribution-invariant statistics in this case.

Thus, Proposition 1 (i) and (ii) provide results for two different schemes of repeated sampling from $\mathcal{S}$, namely, i.i.d. sampling up to a scale factor and pure i.i.d. sampling. Through Proposition 1 (iii) we could consider more general sampling schemes, such as independent sampling from $\mathcal{S}$. To give a practical example, let us consider the spherical context, which corresponds to $\mathcal{S}$ with Y uniformly distributed over S^{n-1}. Then, independent sampling from $\mathcal{S}$ would allow for each Z_i to be independently drawn from a different spherical distribution, *e.g.* Normal, Cauchy, Pearson type II, etc., which does not seem very relevant for practical modelling. Proposition 1 (i) corresponds to each Z_i an independent random vector with the same type of spherical distribution, *e.g.* Normal, but with possibly different scales, whereas Proposition 1 (ii) implies i.i.d. sampling from a particular spherical distribution, such as the standard Normal. In the following Subsection we shall apply these findings to the context of inference robustness for a regression parameter.

3.2. Inference on regression coefficients

We now investigate robustness of sampling theory inference in the model

$$X_i = b_i(\beta) + \sigma_i \varepsilon_i, \quad i = 1, \ldots, p, \quad (3.9)$$

where $X_1, \ldots, X_p$ are n-variate vector observations, $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d. n-dimensional random variables, $\varepsilon_i \stackrel{d}{=} \varepsilon$ (the distribution of which does not depend on β and $\sigma_1, \ldots, \sigma_p$), $\sigma_i > 0$, $i = 1, \ldots, p$, are scale parameters and $b_i(\beta)$, $i = 1, \ldots, p$, are location vectors, parameterized in terms of a common vector $\beta \in \mathcal{B} \subset \Re^m (m \leq n)$ through known functions $b_i(\cdot)$ from $\Re^m$ to $\Re^n$. In the sequel, we shall use the notation $Z_i \equiv X_i - b_i(\beta)$, $i = 1, \ldots, p$, for the error vectors.

Important examples of (3.9) are the standard location-scale model, where $b_i(\beta) = \beta$ with $\beta \in \Re^n$, the case of a common location, where $b_i(\beta) = \beta\iota$ with scalar β and ι an n-dimensional vector of ones, and the regression context, where $b_i(\cdot)$ depends on a matrix of exogenous variables D_i.

Usually, the location or regression parameter β is of interest, whereas the scales are nuisance parameters. We now show how the distribution theory invariance results derived in the previous Subsection can directly be applied in the context of robust classical inference on β. Section 4 will present a parallel Bayesian analysis.

Under the frequentist paradigm, inference on the common parameter β in the model (3.9), will usually be based on some function $t(\cdot)$ of $(Z_1, \ldots, Z_p) = (X_1 - b_1(\beta), \ldots, X_p - b_p(\beta))$. As the distribution of $(Z_1, \ldots, Z_p) = (\sigma_1\varepsilon_1, \ldots, \sigma_p\varepsilon_p)$ does not depend on β, neither will the distribution of $t(Z_1, \ldots, Z_p)$. If, in addition, the distribution of the latter quantity does not depend on $(\sigma_1, \ldots, \sigma_p)$ either, $t(Z_1, \ldots, Z_p)$ is a pivotal quantity, potentially useful in deriving classical inference on β, such as confidence regions and statistics for testing hypotheses, in the presence of nuisance scale parameters.

Since $Z_1, \ldots, Z_p$ are i.i.d. up to a scale factor, Proposition 1 (i) directly leads to

Corollary 1. *Let $X_i = b_i(\beta) + \sigma_i\varepsilon_i$, $i = 1, \ldots, p$, where ε_i are i.i.d. random vectors with $\varepsilon_i \stackrel{d}{=} \varepsilon$ for some $\varepsilon \in \mathcal{S}$ in (2.2), and let $t(\cdot)$ be a measurable function from $\Re^{n\times p}$ to $\Re^k$. Then*

$$\begin{aligned}&t(X_1 - b_1(\beta), \ldots, X_p - b_p(\beta)) \text{ has the same distribution}\\ &\text{for all } \sigma_1, \ldots, \sigma_p > 0 \text{ and for all } \varepsilon \in \mathcal{S},\end{aligned} \tag{3.10}$$

if and only if

$$\begin{aligned}&t(\alpha_1\{X_1 - b_1(\beta)\}, \ldots, \alpha_p\{X_p - b_p(\beta)\}) \stackrel{d}{=} t(X_1 - b_1(\beta), \ldots, X_p - b_p(\beta)),\\ &\text{for all } \alpha_1, \ldots, \alpha_p > 0 \text{ and for all } \varepsilon \in \mathcal{S}.\end{aligned} \tag{3.11}$$

From (3.10) we deduce that the distribution of $t(Z_1, \ldots, Z_p)$ is the same for all $\sigma_1, \ldots, \sigma_p > 0$, and for all $\varepsilon \in \mathcal{S}$. This implies that it is a pivotal quantity that can be used for inference on β in the presence of nuisance scale parameters, and that inferences based on such a $t(Z_1, \ldots, Z_p)$ will be perfectly robust with respect to the choice of ε in the class $\mathcal{S}$. On the other hand, (3.11) tells us that for any $\varepsilon \in \mathcal{S}$ the distribution of $t(Z_1, \ldots, Z_p)$ does not depend on the scaling factors, *i.e.* $t(Z_1, \ldots, Z_p)$ is a pivotal quantity for any $\varepsilon \in \mathcal{S}$. Therefore, Corollary 1 says that any function $t(Z_1, \ldots, Z_p)$ that is a pivotal quantity for any choice of $\varepsilon \in \mathcal{S}$, has the same distribution for all $\varepsilon \in \mathcal{S}$.

Often we possess useful pivotal quantities for the case of one vector observation ($p = 1$). The following construction is a simple fashion to extend them to independent sampling:

Assume that $t_i(\cdot)$ are measurable functions from $\Re^n$ to $\Re^{m_i}$, $i = 1, \ldots, p$, such that $t_i(\alpha Z) \stackrel{d}{=} t_i(Z)$ for all $\alpha > 0$ and for all $Z \in \mathcal{S}$, and that $f(\cdot)$ is any measurable function from $\Re^{m_1 \times \ldots \times m_p}$ to $\Re^k$. Let us define $t : \Re^{n \times p} \mapsto \Re^k$ as $t(z_1, \ldots, z_p) = f\{t_1(z_1), \ldots, t_p(z_p)\}$. Then we obtain

$$
\begin{aligned}
&t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) = f\{t_1(\alpha_1 Z_1), \ldots, t_p(\alpha_p Z_p)\} \stackrel{d}{=} t(Z_1, \ldots, Z_p), \\
&\text{for all } \alpha_1, \ldots, \alpha_p > 0 \text{ and for all } Z_1, \ldots, Z_p \text{ independently drawn from } \mathcal{S}.
\end{aligned}
$$

Thus, any $t(Z_1, \ldots, Z_p)$ constructed in this way is a pivotal quantity that is completely robust under independent sampling from $\mathcal{S}$. Of course, exactly the same pivotal quantity is also distribution-invariant in the more practically interesting situations of i.i.d. sampling up to a scale and pure i.i.d. sampling from $\mathcal{S}$.

In practice, it will often be the case that the given $t_i(\cdot)$'s are scale invariant almost everywhere and not only in distribution, i.e. $t_i(\alpha z) = t_i(z)$ for all $\alpha > 0$ and $z \in \Re^n - A_i$ where $P\{Z \in A_i\} = 0$ for all $i = 1, \ldots, p$ and for all $Z \in \mathcal{S}$. This implies that, for any measurable $f(\cdot)$, $f\{t_1(Z_1), \ldots, t_p(Z_p)\}$ is distribution-invariant not only for $Z_1, \ldots, Z_p$ independently sampled from $\mathcal{S}$, but from the much larger class of multivariate distributions

$$
\mathcal{C} = \{Z : Z = RY \text{ through (2.1), with } Y \text{ distributed as } P\}, \tag{3.12}
$$

characterized by the same marginal probability distribution P for Y as fixed in $\mathcal{S}$, but without imposing independence between R and Y. Thus, the characterizations of robustness derived from Theorem 1 and Proposition 1 are important from the distribution theory point of view, but may often not be required in an applied statistical context.

In the following examples, this simpler case applies, as the pivotal quantities considered will only be functions of the angular coordinates of $Z_1, \ldots, Z_p$.

Example 3.1: spherical linear regression.

Consider the model $X_i = D_i\beta + \sigma_i\varepsilon_i$, $i = 1, \ldots, p$, which is a special case of (3.9) with $b_i(\beta) = D_i\beta$, where D_i is an $n \times m$ fixed matrix of rank m, $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d. with $\varepsilon_i \stackrel{d}{=} \varepsilon$, a random vector with the standard n-variate Normal distribution, $\sigma_i > 0$ is an observation-specific scale parameter, and $\beta \in \Re^m$ is a common vector of m regression parameters.

Let $\hat{\beta}_i$ and $\hat{\sigma}_i$ denote the OLS estimator based on the i-th vector observation and the corresponding estimator of σ_i, respectively. That is, assuming $n > m$,

$$
\hat{\beta}_i = (D_i' D_i)^{-1} D_i' X_i, \quad \hat{\sigma}_i = \left\{ \frac{1}{n-m} (X_i - D_i\hat{\beta}_i)'(X_i - D_i\hat{\beta}_i) \right\}^{1/2}.
$$

The sampling distribution of $S_i = S_i(\beta) \equiv \hat{\sigma}_i^{-1}(\hat{\beta}_i - \beta)$ is an m-variate Student-t distribution with $n - m$ degrees of freedom, location vector 0 and precision matrix $D_i' D_i$, the density function of which is denoted by $f_S^m(s_i | n-m, 0, D_i' D_i)$. Thus the density of $(S_1, \ldots, S_p)$ takes the form $p(s_1, \ldots, s_p | \beta, \sigma_1, \ldots, \sigma_p) = \prod_{i=1}^p f_S^m(s_i | n-m, 0, D_i' D_i)$, which does not depend on the parameters. Therefore, under Normal error vectors with possibly different scale factors, any function of $(S_1, \ldots, S_p)$ is a pivotal quantity, potentially useful for inference on β.

In order to apply our previous results, first note that $S_i = t_i(X_i - D_i\beta)$, where

$$t_i(z) = \left[\frac{1}{n-m} z'\{I_n - D_i(D_i'D_i)^{-1}D_i'\}z\right]^{-1/2} (D_i'D_i)^{-1}D_i'z, \tag{3.13}$$

and $t_i(\alpha z) = t_i(z)$ for all $\alpha > 0$ and for all $z \in \Re^n$ such that $t_i(z)$ is well-defined, *i.e.* for $z \in \Re^n - A_i$, where $A_i = \{D_i\gamma : \gamma \in \Re^m\}$ is an m-dimensional subspace of $\Re^n$. This scale invariance implies that $(S_1, \ldots, S_p)$ is distribution-invariant whenever ε is in a given class $\mathcal{C}$ as in (3.12). Of course, the corresponding marginal distribution on $\mathcal{Y}$ should be chosen such that $P\{\varepsilon \in A_i\} = 0$ for all $i = 1 \ldots, p$, and for all $\varepsilon \in \mathcal{C}$. As an example, the S_i's keep their independent m-variate Student-t distributions when ε ranges in the class $\mathcal{C}$ characterized by the uniform marginal distribution over S^{n-1}. This class $\mathcal{C}$ contains the family of all n-variate spherical distributions as its subclass $\mathcal{S}$ of special interest. The distributional invariance of $S_i = t_i(X_i - D_i\beta)$ when ε_i ranges in a radially decomposable class was also noted in Ng and Fraser (1994).

The aim of this paper is to present conditions which would lead to complete robustness of inference procedures, but we are not proposing any particular robust sampling theory techniques. Therefore, we will not discuss the open question of which functions of $(S_1, \ldots, S_p)$ could be considered and how they should be used for conducting classical inference on β. In the case of a single vector observation, say X_i, some functions of $S_i(\beta)$ are not just arbitrary pivots, but follow from general principles of classical inference. For instance, following the reasoning in Anderson, Fang and Hsu (1986), the likelihood ratio principle for testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \in \Re^m - \{\beta_0\}$ leads to the statistic $F_i(\beta_0) \equiv S_i(\beta_0)'D_i'D_iS_i(\beta_0)/m$ whenever ε_i is in the subclass of spherical distributions where this principle can be applied. Being a function of $S_i(\beta_0)$, this test statistic keeps its $F(m, n-m)$ distribution under the null hypothesis whenever ε_i has a non-Normal spherical distribution (or is in $\mathcal{C}$ characterized by the uniform distribution over S^{n-1}). •

The likelihood ratio principle often leads to robust test statistics and pivotal quantities in the case of one vector observation, as the following Proposition shows. The arbitrariness is then reduced to the particular use we make of those quantities in the case of several vector observations, if we wish to retain robustness.

Proposition 2. *Consider one single observation, X_i, from the model in (3.9) and assume that ε_i follows a continuous υ-spherical distribution with p.d.f. as in (2.3). Then, the likelihood ratio statistic for the test $H_0 : \beta = \beta_0$ versus $H_1 : \beta \in \Re^m - \{\beta_0\}$, provided it exists, takes the form*

$$LR(\beta_0; X_i) = \left[\frac{\inf_{\beta \in \Re^m} \upsilon\{X_i - b_i(\beta)\}}{\upsilon\{X_i - b_i(\beta_0)\}}\right]^n. \tag{3.14}$$

Furthermore, if $b_i(\beta) = D_i\beta$ for some $n \times m$ matrix D_i, $LR(\beta_0; X_i)$ is a scale invariant function of $X_i - D_i\beta_0$. As a consequence, it retains the same distribution under H_0 in the entire υ-spherical class (and even the corresponding class $\mathcal{C}$), for each given choice of $\upsilon(\cdot)$.

Thus, the likelihood ratio statistic takes exactly the same form in the subclass of υ-spherical distributions where it can be computed. From the proof of the Proposition we see that this subclass corresponds to those labelling functions $g(\cdot)$ in (2.3) such that $r^n g(r)$ has a finite supremum in $\Re_+$. From Proposition 2 we also deduce that $LR(\beta_0; X_i)$ retains the same distribution under H_0 in the entire υ-spherical class (and even $\mathcal{C}$) for linear $b_i(\cdot)$.

The following example applies these results to robust inference under l_q-sphericity.

Example 3.2: l_q-spherical observations with common location.

Let us assume the sampling model $X_i = \beta\iota + \sigma_i\varepsilon_i$, $i = 1, \ldots, p$, where $\beta \in \Re$ is the scalar parameter of interest and $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d. with $\varepsilon_i \overset{d}{=} \varepsilon$, an n-dimensional random vector. In this example we will examine robustness when ε follows an l_q-spherical distribution, mentioned in Subsection 2.2, for a fixed value of $q \in (0, \infty)$. The case $q = \infty$ can be analyzed in a similar fashion. A natural reference distribution within the l_q-spherical class corresponds to the elements of ε being i.i.d. draws from an exponential power distribution [see Box and Tiao (1973, chap.3)].

We first consider just one vector observation, $X_i = (X_{1i}, \ldots, X_{ni})'$. The statistic $LR(\beta_0; X_i)$ in (3.14) reduces to

$$LR(\beta_0; X_i) = \left(\sum_{j=1}^{n} |X_{ji} - \tilde{\beta}_i|^q\right)^{n/q} \left(\sum_{j=1}^{n} |X_{ji} - \beta_0|^q\right)^{-n/q}, \qquad (3.15),$$

where $\tilde{\beta}_i$, the maximum likelihood estimator of β based on X_i alone, solves

$$\sum_{j=1}^{n} |X_{ji} - \beta|^{q-1} I_{[0,\infty)}(X_{ji} - \beta) = \sum_{j=1}^{n} |X_{ji} - \beta|^{q-1} I_{(-\infty,0]}(X_{ji} - \beta) \qquad (3.16)$$

and, in particular, is the median of $X_{1i}, \ldots, X_{ni}$ if $q = 1$, and the sample mean, $\overline{X}_i$, if $q = 2$. From Proposition 2, $LR(\beta_0; X_i)$ follows from the likelihood ratio principle for testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$, whenever such principle can be applied, and has the same distribution under H_0 in the entire l_q-spherical class (and the corresponding $\mathcal{C}$).

Thus, in the linear case, the likelihood ratio principle leads to the pivotal quantity $t(X_i - \beta\iota) = LR(\beta; X_i)$, which is distribution-invariant in the class $\mathcal{C}$ of all n-variate distributions of ε_i that share the same marginal distribution of the polar angles as in the reference case of independent sampling from the exponential power distribution. Any measurable function $f\{LR(\beta; X_1), \ldots, LR(\beta; X_p)\}$ will now define a pivotal quantity which is also distribution-invariant, and thus potentially useful for robust inference on β. The choice of $f(\cdot)$ remains an open problem. •

Let us now examine the general nonlinear regression model, where $X_i = b_i(\beta) + \sigma_i\varepsilon_i$. Proposition 2 no longer assures us of the scale invariance of the likelihood ratio statistic $LR(\beta_0; X_i)$ in (3.14). However, robust inference on β can still be conducted on the basis of other statistics which are known to be scale invariant. For instance, the quantity $t_i(X_i - b_i(\beta))$, with $t_i(\cdot)$ as defined in (3.13), still keeps its m-variate Student-t distribution with $n - m$ degrees of freedom, location 0 and precision matrix $D_i'D_i$ for any ε_i in the class

$\mathcal{C}$ corresponding to a uniform distribution over S^{n-1}. As a simple example, the statistic $F_i(\beta_0) = t_i(X_i - b_i(\beta_0))' D_i' D_i t_i(X_i - b_i(\beta_0))/m$ still has an $F(m, n-m)$ distribution when $\beta = \beta_0$ under any spherical distribution on ε_i, and can thus be used for robust testing of $H_0 : \beta = \beta_0$ versus $H_1 : \beta \in \Re^m - \{\beta_0\}$. If D_i is the matrix of first order derivatives of $b_i(\beta)$ (assuming that they exist), evaluated at β_0, then $F_i(\beta_0)$ is the usual F statistic applied to the linearized (around β_0) version of our model. In general, the choice of the particular D_i used in $t_i(\cdot)$ is free, but will certainly affect the efficiency of the inference procedure.

Finally, let us comment on the case of equal scales in (3.9), *i.e.* $\sigma_1 = \ldots = \sigma_p = \sigma$. The random quantities $Z_1, \ldots, Z_p$ are now i.i.d. and Proposition 1 (ii) immediately translates into sufficient and necessary conditions for distribution-invariance of $t(Z_1, \ldots, Z_p)$. In particular, the sufficient condition for distribution-invariance of $t(Z_1, \ldots, Z_p)$ when $\varepsilon \in \mathcal{S}$ is the invariance of its distribution to arbitrary rescaling of each column of $(Z_1, \ldots, Z_p)$ for any $\varepsilon \in \mathcal{S}$. The necessary condition is the invariance to rescaling of the whole matrix. Note that, for the model with common scale σ, this latter condition is equivalent to saying that the distribution of $t(Z_1, \ldots, Z_p)$ does not depend on σ, *i.e.* is a pivotal quantity, for any $\varepsilon \in \mathcal{S}$. However, for $p > 1$ we do not have a characterization of all distribution-invariant pivotal quantities $t(Z_1, \ldots, Z_p)$, and the mere fact that $t(Z_1, \ldots, Z_p)$ is a pivot for any $\varepsilon \in \mathcal{S}$ does not guarantee distribution-invariance.

4. ROBUSTNESS OF BAYESIAN INFERENCE

In this Section we shall consider the issue of robust inference on the regression parameter in the model (3.9) from a Bayesian angle. The nature of the Bayesian paradigm will lead to a somewhat different approach, not based in the distribution theory results of Subsection 3.1, but rather in the joint distribution of observables and parameters.

In Subsection 3.2 we examined robustness of classical inference procedures when ε is a random variable from the class $\mathcal{S}$ in (2.2). As will be clear in the sequel, our Bayesian robustness results will hold in the wider class $\mathcal{C}$, defined in (3.12), where independence between R and Y is not required. The most relevant results are again obtained with observation-specific scales $\sigma_1, \ldots, \sigma_p$, although we shall also briefly comment on the case of equal scale factors. In both situations we will assume a commonly used improper prior density on the scale parameter(s), which conveys the idea of prior ignorance about scale. The following Theorem constitutes our main Bayesian robustness result.

Theorem 2. *Consider the sampling model* $X_i = b_i(\beta) + \sigma_i \varepsilon_i$, $i = 1, \ldots, p$, *where* $\varepsilon_1, \ldots, \varepsilon_p$ *are i.i.d.* n*-dimensional random vectors with* $\varepsilon_i \stackrel{d}{=} \varepsilon$, $\sigma_i > 0$, $i = 1, \ldots, p$, *are scale parameters and* $b_i(\beta)$, $i = 1, \ldots, p$, *are location vectors, parameterized in terms of a common vector* $\beta \in \mathcal{B} \subset \Re^m (m \leq n)$ *through known measurable functions* $b_i(\cdot)$ *from* $\Re^m$ *to* $\Re^n$. *We adopt a prior distribution for* $(\beta, \sigma_1, \ldots, \sigma_p)$ *which is the product measure of the improper density* $p(\sigma_1, \ldots, \sigma_p) = \prod_{i=1}^p p(\sigma_i) = c \prod_{i=1}^p \sigma_i^{-1}$ $(c > 0)$ *and any* σ*-finite prior measure for* β. *Then the joint distribution of* $(X_1, \ldots, X_p, \beta)$ *is exactly the same for all choices of* ε *in the same class* $\mathcal{C}$ *in* (3.12).

From the proof it is easy to see that the key to this result is the invariance of the density $p(\sigma_i) \propto \sigma_i^{-1}$ under the group of transformations $\{\sigma_i r_i : r_i > 0\}$. As a consequence of this, representing $\varepsilon_i = R_i Y_i$ through (2.1), it is obtained that $(\beta, \sigma_1 R_1, Y_1, R_1, \ldots, \sigma_p R_p, Y_p, R_p)$ has the same distribution as $(\beta, \sigma_1, Y_1, R_1, \ldots, \sigma_p, Y_p, R_p)$, from which can be derived that the distribution of $(\beta, \sigma_1 R_1, Y_1, \ldots, \sigma_p R_p, Y_p)$ does not depend on the conditional probability of R_i given Y_i. Thus, the distribution of $(\sigma_1 R_1 Y_1, \ldots, \sigma_p R_p Y_p, \beta) = (X_1 - b_1(\beta), \ldots, X_p - b_p(\beta), \beta)$, and therefore that of $(X_1, \ldots, X_p, \beta)$, do not depend on the conditional probability of R_i given Y_i either. The improper density $p(\sigma_i) \propto \sigma_i^{-1}$ is the only one with this invariance property.

Parallelling the discussion following Theorem 1, the proof of this result does not make use of the one-to-one correspondence between points $z \in \Re^n$ and pairs (y, r). Therefore, Theorem 2 also holds in the more general situation where $\varepsilon_i \stackrel{d}{=} R_i Y_i$ for some positive random variable R_i and we fix the marginal distribution of Y_i. In contrast to Theorem 1, however, independence between R_i and Y_i is now not required.

Theorem 2 says that, under a Jeffreys' type prior for the scale parameters $(\sigma_1, \ldots, \sigma_p)$, the joint distribution of $(X_1, \ldots, X_p, \beta)$ is exactly the same for any choice of $\varepsilon \in \mathcal{C}$. If the marginal distribution of $(X_1, \ldots, X_p)$ is σ-finite, the posterior distribution of β given $(X_1, \ldots, X_p)$ is well-defined, and will clearly also be unaffected by the particular choice of $\varepsilon \in \mathcal{C}$. Thus, posterior inference on β, whenever it can be conducted, is perfectly robust when ε ranges in an entire class $\mathcal{C}$. This result generalizes the finding in Fernández, Osiewalski and Steel (1994), which treated the special case where $p = 1$ (*i.e.* one single vector observation) and all the distributions are dominated by the Lebesgue measure in the corresponding space.

In the case that the distribution of the polar angles $w \in \mathcal{W}$ corresponding to the class $\mathcal{C}$ considered is given through a p.d.f. $p_{\mathcal{W}}(w)$, and if the prior distribution for β is given by some density $p(\beta)$, the resulting distribution for $(X_1, \ldots, X_p, \beta)$ has density function

$$p(x_1, \ldots, x_p, \beta) = cp(\beta) \prod_{i=1}^{p} \frac{\|x_i - b_i(\beta)\|_2^{-n}}{s\left\{h^{-1}\left(\frac{x_i - b_i(\beta)}{\|x_i - b_i(\beta)\|_2}\right)\right\}} p_{\mathcal{W}}\left\{h^{-1}\left(\frac{x_i - b_i(\beta)}{\|x_i - b_i(\beta)\|_2}\right)\right\}, \quad (4.1)$$

where $\{h(w) : w \in \mathcal{W}\} = S^{n-1}$ and $s(w)$ is the factor in the Jacobian of the polar transformation that depends on w. In order to have a proper posterior distribution of β, we require a σ-finite predictive distribution, *i.e.* $p(x_1, \ldots, x_p) = \int_{\mathcal{B}} p(x_1, \ldots, x_p, \beta) d\beta < \infty$ for almost all $(x_1, \ldots, x_p) \in \Re^{n \times p}$. We will now show that for most practically relevant choices of $\mathcal{C}$, the latter condition holds if and only if it holds under Normal sampling.

Proposition 3. *Consider the Bayesian model in Theorem 2, and assume that ε belongs to a class $\mathcal{C}$ corresponding to a p.d.f. $p_{\mathcal{W}}(w)$ for the polar angles $w \in \mathcal{W}$. Then, if there exist two constants K_1 and K_2 such that*

$$0 < K_1 < \frac{p_{\mathcal{W}}(w)}{s(w)} < K_2 < \infty,$$

the marginal distribution of $(X_1, \ldots, X_p)$ is σ-finite if and only if the same holds when ε follows the standard n-variate Normal distribution.

Proposition 3 covers a very wide range of possibilities for $\mathcal{C}$. In particular, if we consider classes $\mathcal{C}$ corresponding to classes $\mathcal{S}$ of υ-spherical distributions, we can see from

$p(w)$ in (2.5) that Proposition 3 applies whenever $\upsilon\{h(w)\}$ is bounded away both from zero and infinity. This covers most of the practically relevant cases such as sphericity, ellipticity, l_q-sphericity or any other υ-spherical class where $\upsilon(\cdot)$ is a norm or where the isodensity contours are bounded and do not touch the origin. Therefore, under weak assumptions on the distribution of the ε_i's, we have established that the marginal distribution of the observables is σ-finite (and thus allows for a proper posterior) if and only if the same holds under Normally distributed errors.

Let us now consider the pure location-scale model (*i.e.* $b_i(\beta) = \beta \in \Re^n$) with a prior density for β, either proper or σ-finite, that does not restrict β to a lower dimensional subspace of $\Re^n$. In this case, we can show that under Normal sampling the predictive density $p(x_1, \ldots, x_p)$ is infinite in a set of positive Lebesgue measure in $\Re^{n\times p}$; thus Bayesian inference is generally precluded, not only under Normality but, applying Proposition 3, under a much wider class of distributional assumptions for ε. Although from Theorem 2 the joint distribution of the observables and the location β is perfectly robust within each class $\mathcal{C}$ for ε, it typically does not allow for posterior inference in the pure location-scale context. Inferentially useful robustness can be obtained, however, when the location is parameterized through a lower dimensional vector, *i.e.* when $\beta \in \mathcal{B} \subset \Re^m$ with $m < n$, as the following example shows.

Example 4.1: spherical linear regression.

Consider the same sampling model as in Example 3.1, *i.e.* p independent Normal regression equations $X_i = D_i\beta + \sigma_i\varepsilon_i$, and the prior $p(\beta, \sigma_1, \ldots, \sigma_p) = cp(\beta)\prod_{i=1}^p \sigma_i^{-1}$.

The resulting Bayesian model was studied by *e.g.* Tiao and Zellner (1964), Dickey (1968), Box and Tiao (1973) and Drèze (1977). For $m < n$, the marginal posterior density of β is proportional to the product of p m-variate Student-t kernels and the prior of β:

$$p(\beta|x_1, \ldots, x_p) \propto p(\beta)\prod_{i=1}^{p} f_S^m(s_i(\beta)|n-m, 0, D_i'D_i),$$

where $S_i(\beta) = \hat{\sigma}_i^{-1}(\hat{\beta}_i - \beta)$ and $\hat{\beta}_i$ and $\hat{\sigma}_i$ are as defined in Example 3.1, $i = 1, \ldots, p$. If $p(\beta)$ is improper uniform over $\Re^m$, this posterior density is proportional to the joint sampling density of $(S_1(\beta), \ldots, S_p(\beta))$ in Example 3.1, considered as a function of β. The latter corresponds to a $p-0$ poly-t distribution [see Drèze (1977)], and is clearly proper for $m < n$. Thus, as a result of the deeper parametric structure on the location vector we can conduct posterior inference on β in this example.

In the Bayesian literature, product form poly-t distributions were obtained as the posteriors resulting from independent Normal samples with different variances. Theorem 2 shows that the assumption of Normality is not necessary and the same marginal posterior $p(\beta|x_1, \ldots, x_p)$ is obtained whenever $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d. random vectors sharing the uniform marginal distribution over S^{n-1}. This implies perfect robustness of Bayesian inference on β in the large class $\mathcal{C}$ of error distributions which includes the class of all n-variate spherical distributions as its subclass of special interest.

From Proposition 3, we also deduce that under any bounded density $p(\beta)$ we can conduct Bayesian inference with most distributional assumptions on ε. Of course, perfectly robust inference is only achieved within each class $\mathcal{C}$ and not across different classes. •

Finally, we remark that in order to obtain the result in Theorem 2, it was crucial to assume p unrelated scale factors σ_i with a Jeffreys' type improper prior on each of them. As the density $p(\sigma_i) \propto \sigma_i^{-1}$ is invariant under the group of transformations $\{\sigma_i r_i : r_i > 0\}$, each scale factor σ_i absorbs the influence of the corresponding R_i. The latter is not possible in the case of equal scales, *i.e.* $X_i = b_i(\beta) + \sigma\varepsilon_i$, $i = 1, \ldots, p$, where $\sigma > 0$ is a common scale parameter. In this case, achieving robustness when ε_i ranges in a class $\mathcal{C}$ of the type described in (3.12) seems impossible whenever $p > 1$.

5. CONCLUSIONS

In this Section we shall summarize the main findings, especially those that we deem of more relevance to the applied modeller, and we shall indicate the parallels that we have found between sampling-theoretic and Bayesian inference.

One of the main aims of this paper was the use of results from multivariate distribution theory in the context of regression models with unknown scale to analyze the robustness of classical inference in certain classes of distributions. Establishing a one-to-one correspondence between points $z \in \Re^n - \{0\}$ and pairs (y, r), where $r > 0$ and y is in some $(n-1)$-dimensional manifold $\mathcal{Y}$, such that $z = ry$, leads to a representation of n-variate random variables Z as $Z = RY$, where R is a positive random variable and Y takes values in $\mathcal{Y}$. This naturally induces classes of multivariate distributions by fixing the marginal distribution of either R or Y. Especially the latter, while keeping R and Y independent, seems of practical interest. In such classes, denoted as $\mathcal{S}$, it is found that functions of a sample of vector observations $t(Z_1, \ldots, Z_p)$, where $Z_1, \ldots, Z_p$ are i.i.d. up to a scale, are distribution-invariant if and only if their distribution is not affected by rescaling each Z_i. Applying this result in the context of the regression model $X_i = b_i(\beta) + \sigma_i\varepsilon_i$, $i = 1, \ldots, p$, where $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d., $\varepsilon_i \stackrel{d}{=} \varepsilon$, we find that a function $t(X_1 - b_1(\beta), \ldots, X_p - b_p(\beta))$ is distribution-invariant for $\varepsilon \in \mathcal{S}$ if and only if it is a pivotal quantity for any $\varepsilon \in \mathcal{S}$. This immediately leads to robust inference on the regression parameter based on such pivots. In addition, we show that for linear $b_i(\cdot)$, the likelihood ratio principle on the basis of one vector observation, X_i, often leads to scale invariant statistics. The only arbitrary element is then the particular use we make of these quantities in the case of repeated sampling.

Another major goal of the paper was to investigate the robustness of Bayesian inference for the same type of model. In addition, we set out to compare these results with the classical findings. For the regression model mentioned above with a prior distribution which is the product measure corresponding to the density $p(\sigma_1, \ldots, \sigma_p) \propto \prod_{i=1}^{p} \sigma_i^{-1}$ and any prior for β, we prove that posterior inference on β (if it is possible) is perfectly robust when ε ranges in a class $\mathcal{C}$ that extends $\mathcal{S}$ by allowing for dependence between R and Y. Although Bayesian robustness holds for this larger class, in practice we shall often be interested in classes of distributions for ε that impose independence between R and Y, thus leading to robustness of both classical and Bayesian inference in the same classes of distributions. In addition, many practical examples of classical invariance can be extended to the class $\mathcal{C}$.

For this model, the key to the classical robustness result is distributional invariance of pivotal quantities with respect to scale transformations, whereas the Bayesian result derives from the invariance with respect to scale transformations of the prior on the scale parameter. In a context of independent sampling, it is crucial for the characterization of robust inference under both paradigms that each vector observation has its own scale parameter. Introducing a regression vector into the model is only important for inference purposes, but does not drive any of the invariance results.

An important issue in the Bayesian paradigm under an improper prior is whether the posterior distribution of the parameter of interest is actually proper, so that it can be used as a basis for inference. We find that for most practically relevant choices of $\mathcal{C}$, properness of the posterior holds in our model if and only if it is achieved in the standard case where ε has an n-variate Normal distribution. In particular, in order to conduct inference on the location, we usually require a reparameterization in terms of a lower dimensional regression parameter.

We feel it is useful, both for practical purposes and for our insight in more fundamental theoretical issues, to amalgamate multivariate distribution theory with sampling-theoretic inference in the context of independent sampling from the commonly used model treated here. In addition, we think our understanding of statistical inference procedures in this model is furthered by contrasting these classical results with their Bayesian counterparts.

APPENDIX

Proof of Theorem 1

First, assume (3.2), which is equivalent to $t(\mathbf{Z}) \stackrel{d}{=} t(\mathbf{Z}^*)$, for all $\mathbf{Z}, \mathbf{Z}^* \in \mathcal{MS}$. From the definition of $\mathcal{MS}$ in (3.1) it follows that if $(Z_1, \ldots, Z_p)$ is in $\mathcal{MS}$, so will be the random variable $(\alpha_1 Z_1, \ldots, \alpha_p Z_p)$ for any $\alpha_1, \ldots, \alpha_p > 0$. Thus, we immediately obtain (3.3).

The proof in the other sense goes as follows, in terms of distribution functions. For any $x \in \Re^k$ we have

$$\begin{aligned} F_{t(Z_1,\ldots,Z_p)}(x) &= F_{t(R_1Y_1,\ldots,R_pY_p)}(x) \\ &= \int_{(0,\infty)^p} F_{t(R_1Y_1,\ldots,R_pY_p)|(R_1,\ldots,R_p)=(r_1,\ldots,r_p)}(x) F_{(R_1,\ldots,R_p)}(dr_1, \ldots, dr_p). \end{aligned}$$

Due to the independence between $(R_1, \ldots, R_p)$ and $(Y_1, \ldots, Y_p)$ assumed in $\mathcal{MS}$, we obtain that $F_{t(R_1Y_1,\ldots,R_pY_p)|(R_1,\ldots,R_p)=(r_1,\ldots,r_p)}(x) = F_{t(r_1Y_1,\ldots,r_pY_p)}(x)$. For any $r_1, \ldots, r_p > 0$, we apply (3.3) with $\alpha_i = r_i$, which leads to $F_{t(r_1Y_1,\ldots,r_pY_p)}(x) = F_{t(Y_1,\ldots,Y_p)}(x)$. Substituting the latter inside the integral, (3.2) follows.

Proof of Proposition 1

From the proof of Theorem 1 we see that (3.2) is equivalent to $t(\alpha_1 Y_1, \ldots, \alpha_p Y_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p)$ for all $\alpha_1, \ldots, \alpha_p > 0$. Using this fact, the proof of Proposition 1 is straightforward.

Proof of Proposition 2

Reparameterizing (β, σ_i) as (β, d_i), where $d_i = \upsilon\{x_i - b_i(\beta)\}/\sigma_i$, the likelihood function takes the form $d_i^n g(d_i)[\upsilon\{x_i - b_i(\beta)\}]^{-n}$, which immediately leads to $LR(\beta_0; X_i)$ in (3.14). We now take $b_i(\beta) = D_i\beta$ and note that

$$LR(\beta_0; X_i) = \left\{ \frac{\inf_{\beta \in \Re^m} \upsilon(X_i - D_i\beta)}{\upsilon(X_i - D_i\beta_0)} \right\}^n = \left\{ \frac{\inf_{\beta \in \Re^m} \upsilon(X_i - D_i\beta_0 - D_i\beta)}{\upsilon(X_i - D_i\beta_0)} \right\}^n,$$

from which it is clear that $LR(\beta_0; X_i)$ is a function of $X_i - D_i\beta_0$. Using the fact that $v(\alpha z) = \alpha v(z)$ for all $\alpha > 0$ and the linearity of $b_i(\cdot)$, we obtain scale invariance of $LR(\beta_0; X_i)$ since

$$\inf_{\beta\in\Re^m} v\{\alpha(X_i - D_i\beta_0) - D_i\beta\} = \alpha \inf_{\beta\in\Re^m} v(X_i - D_i\beta_0 - D_i\beta).$$

Proof of Theorem 2

We shall focus on the case $p = 1$, since the extension to $p > 1$ is immediate. For notational convenience, observation subscripts will be deleted. Representing $\varepsilon = RY$ through (2.1), we shall show that the joint distribution of $(X - b(\beta), \beta)$ does not depend on the distribution of R given Y.

First note that $(X - b(\beta), \beta) = (\sigma\varepsilon, \beta) = (\sigma RY, \beta) = (\lambda Y, \beta)$, where $\lambda = \sigma R$. By assumption, the distribution of (β, σ, Y, R), denoted by $\mathcal{D}_{(\beta,\sigma,Y,R)}$, is the product measure $\mathcal{D}_{(\beta,\sigma,Y,R)} = \mathcal{D}_\beta \times \mathcal{D}_\sigma \times P_{(Y,R)}$, where $\mathcal{D}_\beta$ is the prior measure of β, $\mathcal{D}_\sigma$ corresponds to the density $p(\sigma) = c\sigma^{-1}$ $(c > 0)$, and $P_{(Y,R)}$ is the joint probability measure for (Y, R), which can be decomposed into the marginal probability of Y, P_Y, and the conditional probability of R given Y, $P_{R|Y}$.

If we now consider the one-to-one transformation from (β, σ, Y, R) to (β, λ, Y, R), where $\lambda = \sigma R$, the distribution of (β, λ, Y, R) will be given by $\mathcal{D}_{(\beta,\lambda,Y,R)} = \mathcal{D}_\beta \times \mathcal{D}_{(\lambda,Y,R)}$, where the distribution of (λ, Y, R) is computed in the following way:
For each measurable set C, $\mathcal{D}_{(\lambda,Y,R)}(C) = \mathcal{D}_{(\sigma,Y,R)}\{f^{-1}(C)\}$, where $f^{-1}(C) = \{(\sigma, y, r) : (\sigma r, y, r) \in C\}$. Applying the Classical Product Measure Theorem, we obtain

$$\mathcal{D}_{(\sigma,Y,R)}\{f^{-1}(C)\} = \int_{\mathcal{Y}\times(0,\infty)} \mathcal{D}_\sigma\{f^{-1}(C)_{(y,r)}\} P_{(Y,R)}(dy, dr),$$

where $f^{-1}(C)_{(y,r)} = \{\sigma : (\sigma, y, r) \in f^{-1}(C)\} = \{\sigma : (\sigma r, y, r) \in C\} = \{\sigma : \sigma r \in C_{(y,r)}\}$, with $C_{(y,r)} = \{\sigma : (\sigma, y, r) \in C\}$. The invariance property of $\mathcal{D}_\sigma$ under the group of transformations $\{r\sigma : r > 0\}$, implies that for each value of $r > 0$, $\mathcal{D}_\sigma\{f^{-1}(C)_{(y,r)}\} = \mathcal{D}_\sigma\left\{\sigma : \sigma r \in C_{(y,r)}\right\} = \mathcal{D}_\sigma\{C_{(y,r)}\}$. Substituting the latter expression inside the integral above leads to $\mathcal{D}_{(\lambda,Y,R)}(C) = \mathcal{D}_{(\sigma,Y,R)}\{f^{-1}(C)\} = \mathcal{D}_{(\sigma,Y,R)}(C)$, and thus $\mathcal{D}_{(\beta,\lambda,Y,R)} = \mathcal{D}_\beta \times \mathcal{D}_\lambda \times P_{(Y,R)}$, where $\mathcal{D}_\lambda = \mathcal{D}_\sigma$.

In order to recuperate the distribution of $(X - b(\beta), \beta) = (\lambda Y, \beta)$, we look at the distribution of (β, λ, Y). By definition of the marginal distribution, the measure of any set A is given by $\mathcal{D}_{(\beta,\lambda,Y)}(A) = \mathcal{D}_{(\beta,\lambda,Y,R)}\{A \times (0,\infty)\} = (\mathcal{D}_\beta \times \mathcal{D}_\sigma \times P_Y)\{A\} P_{R|Y}\{(0,\infty)\}$. As $P_{R|Y}\{(0,\infty)\} = 1$ for all possible choices of probability measures $P_{R|Y}$, the distribution of (β, λ, Y) does not depend on $P_{R|Y}$. Therefore, the distributions of $(\lambda Y, \beta) = (X - b(\beta), \beta)$ and (X, β) do not depend on $P_{R|Y}$ either.

Proof of Proposition 3

A Normal distribution for ε corresponds to $p_{\mathcal{W}}(w) \propto s(w)$ which, combined with (4.1), leads to the result.

REFERENCES

Anderson, T.W., Fang, K.T. and Hsu, H. (1986), "Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions," *The Canadian Journal of Statistics*, 14, 55-59.

Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis,* New York: Springer-Verlag.

Box, G.E.P. and Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis,* Reading: Addison-Wesley.

Dawid, A.P. (1977), "Spherical matrix distributions and a multivariate model," *Journal of the Royal Statistical Society B*, 39, 254-261.

Dickey, J.M. (1968), "Three multidimensional-integral identities with Bayesian applications," *Annals of Mathematical Statistics*, 39, 1615-1628.

Drèze, J.H. (1977), "Bayesian regression analysis using poly-t densities," *Journal of Econometrics*, 6, 329-354.

Fang, K.T., Kotz, S. and Ng, K.W. (1990), *Symmetric Multivariate and Related Distributions,* London: Chapman and Hall.

Fang, K.T. and Zhang, Y.T. (1990), *Generalized Multivariate Analysis,* Berlin: Springer-Verlag.

Fernández, C., Osiewalski, J. and Steel, M.F.J. (1994) "The continuous multivariate location-scale model revisited: a tale of robustness," *Biometrika*, 81, 588-594.

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

Kelker, D. (1970), "Distribution theory of spherical distributions and a location-scale generalization," *Sankhyā* A, 32, 419-430.

Nachtsheim, C.J. and Johnson, M.E. (1988), "A new family of multivariate distributions with applications to Monte Carlo studies," *Journal of the American Statistical Association,* 83, 984-989.

Ng, K.W. and Fraser, D.A.S. (1994), "Inference for linear models with radially decomposable error," in *Multivariate Analysis and its Applications*, IMS Lecture Notes-Monograph Series, Vol. 24, 359-367.

Osiewalski, J. and Steel, M.F.J. (1993), "Robust Bayesian inference in l_q-spherical models," *Biometrika,* 80, 456-460.

Tiao, G.C. and Zellner, A. (1964), "Bayes' theorem and the use of prior knowledge in regression analysis," *Biometrika,* 51, 219-230.