Inference robustness in multivariate models with a scale parameter

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

Extended Abstract

We formulate a general representation of points $z \in \Re^n - \{0\}$ in terms of pairs (y, r), where $r > 0$, y lies in some space $\mathcal{Y}$, and $z = ry$. In addition, we impose that the representation is unique. An example of such a representation is polar coordinates, which corresponds to $\mathcal{Y} = S^{n-1}$, $r = \|z\|_2$, the Euclidean norm, and $y = z/\|z\|_2$, a point in the unit sphere S^{n-1}.

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 interesting families of multivariate distributions can be interpreted in this unifying framework. For instance, the well-known spherical class corresponds to choosing $\mathcal{Y} = S^{n-1}$ and Y uniformly distributed over S^{n-1}, whereas the anisotropic class is obtained by taking $\mathcal{Y} = S^{n-1}$ and fixing an $\sqrt{\chi^2_n}$ distribution for the Euclidean radius R. Other families of multivariate distributions, such as l_q-spherical and υ-spherical classes, are also seen to correspond to certain choices of $\mathcal{Y}$ and distributions of Y.

Some classical inference procedures can be shown to be completely robust or distribution free in these classes of multivariate distributions, generated by fixing the distribution of either R or Y, while imposing independence between them. These findings are used in the practically relevant contexts of location-scale and pure scale models, and we explicitly distinguish the case of sampling one multivariate observation from the inferentially more useful situation of independent sampling from multivariate distributions. Finally, we present a robust Bayesian analysis for the same models 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. Some examples using spherical, l_q-spherical and l_q-anisotropic sampling distributions are presented.

Inference Robustness in Multivariate Models With a Scale Parameter

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

ABSTRACT

We formulate a general representation of points $z \in \Re^n - \{0\}$ in terms of pairs (y, r), where $r > 0$, y lies in some space $\mathcal{Y}$, and $z = ry$. In addition, we impose that the representation is unique. An example of such a representation is polar coordinates.

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, 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 contexts of location-scale and pure scale models. Finally, we present a robust Bayesian analysis for the same models 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-freeness; Scale invariance; Pivotal quantity; Bayesian inference; Regression model.

Carmen Fernández is Research Associate at the Institut de Statistique, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium. 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, during the work on this paper. Mark F.J. Steel is Senior Research Fellow at the CentER for Economic Research and Associate Professor at the Department of Econometrics, Tilburg University, 5000 LE Tilburg, The Netherlands. The first author acknowledges support from D.G.I.C.Y.T. under grant number PB91-0014, 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, and seminar participants at the Institut de Statistique, Louvain-la-Neuve, 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-free or 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 in Section 3 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. In addition, Section 3 extends certain classes of continuous distributions to this more general context, and investigates some of their properties in more detail.

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. Section 4 presents results characterizing distribution-free functions of matrix random variables, and applies these findings to robust classical inference in the context of regression models with scale and pure scale models. A parallel robust Bayesian analysis of these models is described in Section 5. For both paradigms, the robustness results essentially hinge upon the presence of a scale parameter in the model.

Whereas the classical results are derived from distribution theory in a rather general context, we find that practical examples are often illustrative of a simpler special case. In addition, our classical findings pertain to the distributional properties of certain pivots, but do not suggest any particular way of using these pivots in inferential procedures. In contrast, the Bayesian robustness results are immediately applicable to practical examples and inference procedures.

The final Section groups some conclusions and puts the similarities between both classical and Bayesian approaches in perspective.

2. A REPRESENTATION OF $\Re^n$

In this Section, 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 certain useful classes of distributions in which robustness results can be derived.

We shall represent each point $z \in \Re^n - \{0\}$ through a pair (y, r), where $r > 0$, y is in some set $\mathcal{Y}$, and $z = ry$. In addition, we impose that the representation is unique, which means that for each point z there exists a unique pair (y, r) and viceversa. This implies that $\mathcal{Y}$ is an $(n-1)$-dimensional manifold, and can be represented in terms of $w \in \mathcal{W} \subset \Re^{n-1}$ through a one-to-one function $k(\cdot)$, i.e. $\mathcal{Y} = \{k(w) : w \in \mathcal{W}\}$. Without loss of generality, we shall characterize $\mathcal{Y}$ through angular polar coordinates. 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 $\mathcal{Y}$ 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 or l_2-radius. This leads to the usual polar representation. Another possibility would be to take $\mathcal{Y} = \{x \in \Re^n : \|x\|_q \equiv (\sum_{i=1}^n |x_i|^q)^{1/q} = 1\}$, the unit l_q-sphere, for some $q \in [1, \infty) - \{2\}$, in which case $r = \|z\|_q$ describes the l_q-norm or l_q-radius, whereas $y = z/\|z\|_q$ is the corresponding point on the unit l_q-sphere.

The representation described here 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 $Y = k(W)$, and we can alternatively represent Z as

$$Z = Rk(W). \tag{2.2}$$

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). In the sequel, we shall either use (2.1) or (2.2), whichever is more convenient. Clearly, any such Z can always be characterized in terms of $Y = Z/\|Z\|_2$, which takes values on S^{n-1}, and $R = \|Z\|_2$. However, as we shall explain in the following, this standard polar representation does not suffice for our purposes.

In particular, the correspondence between the distributions of Z and (Y, R) naturally leads to the definition of classes of distributions characterized by fixing the marginal distribution of either R or Y. Many well-known families of multivariate distributions can be generated in this way.

If we consider the class generated through fixing the marginal distribution of Y, while allowing for any conditional distribution of $R|Y$, the standard polar representation will be enough as there is a one-to-one correspondence between S^{n-1} and any $\mathcal{Y}$. Therefore, fixing a distribution on S^{n-1} uniquely identifies a distribution on any other $\mathcal{Y}$, and any choice of $\mathcal{Y}$ in the representation in (2.1) would induce the same class of distributions. In fixing the marginal distribution of R, on the other hand, the particular choice of $\mathcal{Y}$ is crucial, as there is no one-to-one correspondence between different representations of the radius if Y is not specified. For example, the class generated by a certain distribution for the Euclidean radius will not coincide with the class induced by fixing the distribution of the l_1-radius.

Therefore, we need more than the standard representation if we wish to generate such classes. In addition, the most interesting classes will be seen to correspond to independence between R and Y, while keeping the distribution of either R or Y fixed. In this case, the particular choice of $\mathcal{Y}$ in (2.1) will always matter. For instance, independence between the Euclidean radius and W (and thus Y in S^{n-1}), is not equivalent to independence between the l_1-radius and W (and thus Y in the unit l_1-sphere). Therefore, if we impose independence between R and Y, we always need our more general framework.

The next Section will discuss some classes of multivariate distributions naturally obtained from our representation under independence between R and Y.

3. CLASSES OF MULTIVARIATE DISTRIBUTIONS

In this Section, we shall define classes of probability distributions on $\Re^n - \{0\}$ through the representation in (2.1), that, for a given $\mathcal{Y}$, correspond to a particular choice of the distribution of either Y or R, while imposing independence between them.

3.1. FIXING THE DISTRIBUTION OF Y

Given a choice of $\mathcal{Y}$ and any probability distribution P_2 on $\mathcal{Y}$, we define the class $\mathcal{S}$ as the following set of random variables on $\Re^n - \{0\}$

$$\mathcal{S} = \{Z : Z = RY \text{ through (2.1), with } R \text{ and } Y \text{ independent and } Y \text{ distributed as } P_2\}. \tag{3.1}$$

We now consider the class of distributions corresponding to the random variables in $\mathcal{S}$. By varying the choice of $\mathcal{Y}$ and P_2, we generate many useful classes of multivariate distributions. We shall now show that many classes of distributions that have appeared in the literature can be interpreted in the framework of (3.1). In addition, our representation in Section 2 sheds new light on the properties of some of these classes. In particular, we can mention:

3.1.1 Spherical Distributions

This class has a long-standing tradition in multivariate distribution theory and is well-documented in e.g. Kelker (1970) and Fang *et al.* (1990).

Originally, sphericity was defined in terms of distributional invariance under orthogonal transformations, i.e. Z has a spherical distribution if for every Γ belonging to the group of $n \times n$ orthogonal matrices, denoted by $\mathcal{O}(n)$, $\Gamma Z \stackrel{d}{=} Z$, i.e. ΓZ and Z have the same distribution. However, it is well-known [see Theorem 2.5 of Fang *et al.* (1990)] that the spherical class can alternatively be characterized in terms of $\mathcal{S}$ in (3.1) by choosing $\mathcal{Y}$ to be the unit sphere S^{n-1}, and Y uniformly distributed over S^{n-1}.

In the case of continuity, these distributions are characterized by spherical isodensity sets, whereas the labelling function can be chosen freely (provided, of course, that the resulting density function integrates to unity). Important continuous special cases of

the spherical class are Normal, Student-t and Pearson Type II distributions. Normality constitutes the natural reference case in this class.

3.1.2 l_q-Spherical Distributions

For the continuous case, these distributions were defined in Osiewalski and Steel (1993), through properties of the density function. Whereas they use a location-scale model as they focus on inference, we shall, in this Subsection, present l_q-sphericity without location and scale parameters. In particular, the density function of $Z = (Z_1, \ldots, Z_n)'$ in $\Re^n$ was chosen to be

$$p(z) = g\{v_q(z)\} \tag{3.2}$$

where

$$v_q(z) = \begin{cases} \left(\sum_{i=1}^n |z_i|^q\right)^{1/q} & \text{if } 0 < q < \infty \\ \max_{i=1,\ldots,n} |z_i| & \text{if } q = \infty \end{cases}$$

and $g(\cdot)$ is any nonnegative function such that $p(z)$ is a proper density.

Given q, the l_q-spherical class is generated by allowing for all possible choices of the labelling function $g(\cdot)$. Note that $v_q(\cdot)$ corresponds to the l_q-norm extended to values of $q \in (0, 1)$. Thus, the class generated by (3.2) corresponds to all continuous multivariate distributions with l_q-spheres as isodensity sets for a particular choice of q. We shall now verify that, for any finite q, this class also fits in the framework of (3.1). For $q = \infty$ a similar derivation can be given.

Indeed, if we define $R = v_q(Z)$ and $D_i = (|Z_i|/R)^q$, $i = 1, \ldots, n$, Osiewalski and Steel (1993) show that (3.2) leads to a product of an n-variate Dirichlet distribution for $D = (D_1, \ldots, D_n)'$ with parameters $1/q$ and a distribution for $R|D$ with probability density function

$$p(r|d) = p(r) = \frac{[2\Gamma\{1 + (1/q)\}]^n}{\Gamma\{1 + (n/q)\}} n r^{n-1} g(r).$$

Following Dickey and Chen (1985), they use the stochastic representation

$$Z \stackrel{d}{=} R(\Delta \times D^{1/q}),$$

where R, Δ and D are all independent, $D^{1/q}$ denotes a coordinatewise power, $\Delta \times D^{1/q}$ is a coordinatewise product of vectors, and the n elements of Δ independently take the value 1 or -1 with probability 1/2. Defining $Y = Z/R$, we are back in the representation $Z = RY$ as in (2.1), with $\mathcal{Y}$ the unit l_q-sphere. Furthermore, $Y \stackrel{d}{=} \Delta \times D^{1/q}$, and thus has a fixed distribution for given q, whereas, by leaving $g(\cdot)$ free, we have allowed for any continuous probability distribution for R, always imposing independence between R and Y. Thus, extending the class of continuous l_q-spherical distributions to possibly noncontinuous distributions of R, we obtain a class that can be described by (3.1).

A natural reference case for the l_q-spherical class is generated by assuming that the elements of Z are independently sampled from exponential power distributions [Box and Tiao, 1973, Ch. 3].

3.1.3 υ-Spherical Distributions

Pursueing the idea of characterizing classes of continuous distributions through their isodensity sets, Fernández *et al.* (1995) consider more flexible shapes. For inference purposes they use a location-scale context, which shall not be explicited in this Subsection. They introduce continuous υ-spherical distributions through the following density function of $Z = (Z_1, \ldots, Z_n)'$ in $\Re^n$:

$$p(z) = g\{\upsilon(z)\}, \tag{3.3}$$

where $\upsilon(\cdot)$ is a scalar function such that
(i) $\upsilon(\cdot) > 0$, except possibly on a set of Lebesgue measure zero,
(ii) $\upsilon(\alpha z) = \alpha\upsilon(z)$, for all $\alpha \geq 0$, $z \in \Re^n$,
and $g(\cdot)$ is a nonnegative labelling function. Furthermore, $\upsilon(\cdot)$ and $g(\cdot)$ are such that (3.3) 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. Clearly, continuous spherical and l_q-spherical distributions correspond to continuous υ-spherical distributions with $\upsilon(\cdot)$ the l_2-norm and the l_q-norm, respectively. The name "υ-spherical" is motivated by the fact that all members of the same class have common isodensity sets of the form:

$$\{z \in \Re^n : \upsilon(z) = \beta\} \tag{3.4}$$

where β is a positive constant. We shall call the set in (3.4) the "υ-sphere" of "υ-radius" β.

We will now show that the υ-spherical class can also be interpreted in the framework of (3.1), which facilitates the derivation of certain novel properties.

Representing $\Re^n$ in terms of polar coordinates (w, u), where $w \in \mathcal{W} \subset \Re^{n-1}$ denotes the angular coordinates and $u > 0$ is the Euclidean radius, we obtain

$$z = uh(w), \quad \text{for all } z \in \Re^n - \{0\}, \tag{3.5}$$

where $h(w)$ lies in the unit sphere S^{n-1}. Now, (3.3) leads to

$$p(w, u) = g[\upsilon\{uh(w)\}]u^{n-1}s(w), \tag{3.6}$$

where $u^{n-1}s(w)$ is the appropriate Jacobian.

However, the correspondence in (3.5) does not lead to a representation of Z in terms of a product of independent random quantities, as is obvious from (3.6). Nevertheless, if we consider a further transformation from (w, u) to (w, r), with $r = \upsilon\{uh(w)\} = \upsilon(z)$, the υ-radius from (3.4), it is easy to see that

$$p(w, r) = s(w)[\upsilon\{h(w)\}]^{-n}r^{n-1}g(r), \tag{3.7}$$

from which independence between $W = h^{-1}(Z/\|Z\|_2)$ and $R = \upsilon(Z)$ is immediately deduced. Thus, we have a representation of the form (2.2), with $k(w) = h(w)/\upsilon\{h(w)\}$ whenever $\upsilon\{h(w)\} > 0$ and $ch(w)$, where $c > 0$ is an arbitrary constant, otherwise. Note

that $\mathcal{Y} = \{k(w) : w \in \mathcal{W}\} = \{x \in \Re^n : v(x) = 1\} \cup \{ch(w) : v\{h(w)\} = 0\}$ and, therefore, represents the union of the unit v-sphere with some other set corresponding to a set of polar angles with Lebesgue measure (in $\Re^{n-1}$) zero. In addition, from (3.7) we see that only $v(\cdot)$ determines the distribution of W (or, equivalently, the distribution on $\mathcal{Y}$), which concentrates all the mass on the unit v-sphere and is characterized by the density function

$$f_2(w) \propto s(w)[v\{h(w)\}]^{-n}. \tag{3.8}$$

The distribution of R, on the other hand, is entirely determined by $g(\cdot)$, with density function

$$p(r|w) = p(r) \propto r^{n-1} g(r). \tag{3.9}$$

Furthermore, note that the only restriction on $g(\cdot)$ is given through properness of (3.3), and thus of (3.9). This implies, from (3.9), that we can accommodate any proper density for R. Thus, for a given $v(\cdot)$, the class of continuous v-spherical distributions corresponds to a subset of a particular $\mathcal{S}$ in (3.1), where Y takes values on the unit v-sphere with its distribution fixed by the choice of $v(\cdot)$ through (3.8). If we extend the class of continuous v-spherical distributions by also allowing for any noncontinuous distribution of R, we cover the entire class $\mathcal{S}$.

Thus, for any choice of $v(\cdot)$, the v-spherical class consists of all distributions in $\Re^n$ that share the common marginal distribution over the unit v-sphere corresponding to the density of the polar angles in (3.8), while imposing independence between the polar angles and the v-radius. Note that the integrability condition on (3.3) does not restrict the class of density functions $f_2(w)$ we can accommodate. For any proper $f_2(w)$, we can always find a function $v(\cdot)$ such that the induced v-spherical distributions are characterized by the marginal density $f_2(w)$ for the polar angles. In particular, we first define $v(\cdot)$ on S^{n-1} as

$$v\{h(w)\} \propto \{s(w)/f_2(w)\}^{1/n} \tag{3.10}$$

and extend it to all of $\Re^n$ by using property (ii) of the definition in (3.3), i.e.

$$v(z) = v\{uh(w)\} = uv\{h(w)\}. \tag{3.11}$$

The density $f_2(w)$ in (3.8) uniquely defines a probability distribution on the unit v-sphere. In the special case of sphericity where $v(\cdot)$ is the Euclidean norm, we obtain $f_2(w) \propto s(w)$ which implies a uniform distribution on the unit sphere S^{n-1}. If $v(\cdot)$ is any other function, we lose the uniformity on S^{n-1} and a natural question to ask is whether we recuperate this uniformity on the corresponding unit v-sphere instead. The answer, in general, is no. Even for l_q-spherical distributions with $n = 2$, one can prove that such uniformity only holds for $q = 1, 2$ or ∞. In fact, uniformity on the l_q-sphere can, in general, not even be salvaged if we consider the l_q-distance instead of the Euclidean distance. So, barring some special cases, $f_2(w)$ can not be attributed any natural interpretation in terms of uniformity on the unit v-sphere. However, the distribution induced by $f_2(w)$ on the unit v-sphere admits an alternative interpretation:

Let A_v be any measurable set on the unit v-sphere. Then the corresponding polar angles constitute the following Borel set in $\mathcal{W}$:

$$A = \{h^{-1}(z/\|z\|_2) : z \in A_v\}.$$

Alternatively, we can write

$$A_v = \{z \in \Re^n : h^{-1}(z/\|z\|_2) \in A \text{ and } v(z) = 1\}$$

and we can show

$$P(A_v) = \int_A f_2(w)dw \propto \int_B dz,$$

where

$$B = \{z \in \Re^n : h^{-1}(z/\|z\|_2) \in A \text{ and } v(z) \leq 1\}.$$

Thus, the probability of any measurable set on the unit v-sphere is proportional to the hypervolume that it generates inside the v-sphere. This fact immediately explains why the interpretation of a uniform distribution on the v-sphere generally does not apply. Both interpretations only coincide in very special cases, such as sphericity and l_1-sphericity.

We stress, once more, the great flexibility of the v-spherical class. As mentioned above, specific choices of $v(\cdot)$ generate the spherical and l_q-spherical distributions. Another interesting example of a v-spherical class are the elliptical distributions, with covariance structure V, which correspond to $v(z) \propto (z'V^{-1}z)^{1/2}$ for each positive definite symmetric matrix V of dimension $n \times n$.

As an example, consider the issue of planetary motion in astronomy. Kepler's First Law states that planets describe an elliptical orbit, with the sun located at one focal point. According to Kepler's Second Law, the probability of finding the planet (at any given time) in a particular subset of its orbit is proportional to the area under that subset with respect to the sun. Thus, from our previous discussion, a planet's location has a v-spherical distribution in the plane containing its orbit, with the v-sphere defined by the orbit and a Dirac distribution on the corresponding v-radius. Note, however, that this does not imply an elliptical distribution as defined above, since planetary orbits are not symmetric around the sun, which is in one focal point, and not in the center of the ellipse.

All classes of distributions mentioned above share the appealing property that in the continuous case all isodensity sets have a common shape, whereas the choice of the labelling function is kept entirely free. Thus, these classes are particularly suited to modelling where one often has a much better idea of the shape of the isodensity sets than of the labelling function (e.g. tail behaviour). Through judicious choices of $v(\cdot)$ in the, most general, v-spherical class, we can accommodate a wide range of possible isodensity surfaces [see Fernández *et al.* (1995)]. Note that the assumption of independence between the distribution on the unit v-sphere and the v-radius plays a crucial role in obtaining this characteristic. Consider the class of all continuous random variables Z in $\Re^n$ which share a common density $f_2(w)$ of the polar angles. Choosing $v(\cdot)$ to be compatible with $f_2(w)$ through (3.10) and (3.11), we can easily obtain that the density function of Z can be expressed as

$$p(z) \propto p(r|w)r^{1-n},$$

where $r = v(z)$ is the v-radius. Obviously, imposing independence between R and W will assure us that all isodensity sets are v-spheres as defined in (3.4). As an example, consider

the class of all continuous distributions in $\Re^2$ which correspond to the, so-called, cardioid distribution for the polar angle, characterized by

$$f_2(w) = \frac{1 + 0.6\cos(w)}{2\pi} I_{[0,2\pi)}(w).$$

Fernández *et al.* (1994) present some members of this class, without, however, imposing independence between W and the corresponding v-radius for $z = (z_1, z_2)'$,

$$v(z) \propto \|z\|_2^{3/2}(0.6z_2 + \|z\|_2)^{-1/2},$$

derived through (3.10) and (3.11). In two of their examples they impose independence between W and the l_2-radius instead, thus generating two members of a class $\mathcal{S}$ which is not v-spherical. The resulting density functions possess very different isodensity sets for each choice of the conditional density of the v-radius. Alternatively, if we focus on the subclass induced by independence between the v-radius and W, we obtain all the possible densities with isodensity contours as displayed in Figure 1. Thus, this v-spherical subclass seems the most interesting one from a modelling perspective.

3.2 FIXING THE DISTRIBUTION OF R

In this Subsection, we shall be concerned with classes of multivariate distributions which are generated from the representation in (2.1) by choosing a particular distribution for R, while retaining the independence between R and Y.

Given a particular choice of $\mathcal{Y}$, and a fixed probability distribution P_1 on $\Re_+$, we define the class of random variables

$$\mathcal{R} = \{Z : Z = RY \text{ through (2.1), with } R \text{ and } Y \text{ independent and } R \text{ distributed as } P_1\}. \tag{3.12}$$

Classes of distributions generated through choosing $\mathcal{Y}$ and P_1 in (3.12) have received much less attention in the literature than their conterparts based on $\mathcal{S}$ in (3.1). Let us mention the following example.

3.2.1 Anisotropic Distributions

Nachtsheim and Johnson (1988) introduce the anisotropic family of multivariate distributions, which exactly fits into the framework of (3.12), 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. However, 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. Nachtsheim and Johnson (1988) conduct a simulation study

to investigate the robustness properties of Hotelling's T^2 statistic under independent and identically distributed (i.i.d.) sampling from distributions in the anisotropic class.

Similarly, we can start from a situation where the elements of Z are independently sampled from an exponential power distribution. If we then fix the distribution of $Y = Z/v_q(Z)$ and leave that of the l_q-radius $R = v_q(Z)$ free, but independent of Y, we generate an l_q-spherical class as discussed in Subsection 3.1.2. Conversely, keeping R independent of Y and distributed as $(\chi^2_{2n/q})^{1/q}$ for finite values of q and as Beta$(n, 1)$ for $q = \infty$ [see Osiewalski and Steel (1993)], and letting the distribution of Y change, we will define the class of l_q-**anisotropic** distributions, for any given $q \in (0, \infty]$.

Having established the practical use of the classes generated through $\mathcal{S}$ in (3.1) and $\mathcal{R}$ in (3.12), we shall now focus on robustness of inference procedures within these classes.

4. 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 (3.1). Therefore, most of the discussion in this Section will be devoted to classes $\mathcal{S}$. Subsection 4.1 derives results on distribution-free functions of random variables for a matricvariate generalization of $\mathcal{S}$, and Subsection 4.2 applies this theory to robust inference on location in a regression model in a context of multiple observations from $\mathcal{S}$. For a matricvariate generalization of the classes $\mathcal{R}$ in (3.12), Subsection 4.3 will group both theory and inference results.

4.1 DISTRIBUTION INVARIANCE RESULTS IN $\mathcal{S}$

In practice, inference will typically be conducted on certain parameters in the model on the basis of more than one vector observation. We do not yet introduce parameters into the sampling model at this stage (we reserve that until the next Subsection), but we shall now consider matrices of observables, rather than vectors.

In particular, let us now focus on matrix random variables $Z = (Z_1, \ldots, Z_p)$ 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 $\mathcal{Y}_i$. Then we define

$$\mathcal{MS} = \{Z = (Z_1, \ldots, Z_p) : Z = (R_1 Y_1, \ldots, R_p Y_p) \text{ with } (R_1, \ldots, R_p) \text{ independent of } (Y_1, \ldots, Y_p) \text{ and the distribution of } (Y_1, \ldots, Y_p) \text{ fixed}\}. \tag{4.1}$$

Thus, for a given $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$ and a fixed distribution of $(Y_1, \ldots, Y_p)$, the class $\mathcal{MS}$ is generated by considering all possible distributions for $(R_1, \ldots, R_p)$, while imposing independence between $(R_1, \ldots, R_p)$ and $(Y_1, \ldots, Y_p)$. Clearly, in the case $p = 1$, the class $\mathcal{MS}$ reduces to our earlier $\mathcal{S}$ for vector distributions in (3.1). Also note that if $\mathcal{Y}_i = \mathcal{Y}$ for all $i = 1, \ldots, p$, and the fixed joint distribution of $(Y_1, \ldots, Y_p)$ leads to the same marginal distribution for each Y_i on $\mathcal{Y}$, then each of the components of Z belongs to the same class

$\mathcal{S}$. A class in (4.1) that has appeared in the literature is generated by choosing $\mathcal{Y}_i = S^{n-1}$, $i = 1, \ldots, p$, and fixing $Y_1, \ldots, Y_p$ 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 $p = 1$.

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

Theorem 1. *Let $\mathcal{MS}$ be the class as defined in (4.1) and $t(\cdot)$ be a measurable function from $\Re^{n\times p}$ to $\Re^k$. Then*

$$t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p), \text{ for all } (Z_1, \ldots, Z_p) \in \mathcal{MS} \tag{4.2}$$

if and only if

$$t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p), \text{ for all } \alpha_1, \ldots, \alpha_p > 0, \text{ for all } (Z_1, \ldots, Z_p) \in \mathcal{MS}. \tag{4.3}$$

Proof: see Appendix. •

While we present (4.2) 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 (4.2) as

$$t(Z) \stackrel{d}{=} t(Z^*), \text{ for all } Z, Z^* \in \mathcal{MS}.$$

Thus, this condition is equivalent to saying that $t(Z)$ is distribution-free in $\mathcal{MS}$. On the other hand, (4.3) can be interpreted as invariance of the distribution of $t(Z)$ with respect to changes of scale for each of the columns of Z. In fact, Theorem 1 tells us that the latter property exactly characterizes all the functions of Z that are distribution-free in $\mathcal{MS}$.

In proving that (4.3) implies (4.2) (Appendix), we only use (4.3) to obtain that

$$t(\alpha_1 Y_1, \ldots, \alpha_p Y_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p), \text{ for all } \alpha_1, \ldots, \alpha_p > 0, \tag{4.4}$$

which looks weaker than (4.3) but can actually be shown to be equivalent. Thus, we can, equivalently, state Theorem 1 as follows: $t(Z)$ is distribution-free in $\mathcal{MS}$ if and only if it is distribution-free in the subset of $\mathcal{MS}$ corresponding to Dirac distributions on $(\alpha_1, \ldots, \alpha_p)$ for $(R_1, \ldots, R_p)$, for all $\alpha_1, \ldots, \alpha_p > 0$. Note that the set of all distributions for $(R_1, \ldots, R_p)$ on $(0, \infty)^p$ is the convex hull of the Dirac distributions described above. Using this fact and the independence between $(R_1, \ldots, R_p)$ and $(Y_1, \ldots, Y_p)$, we can directly derive that $\mathcal{MS}$ is the convex hull of its subset corresponding to all Dirac distributions for $(R_1, \ldots, R_p)$. The latter result carries over to distributions of $t(Z)$ and provides us with an alternative way to prove and interpret the Theorem. This sheds some light on why the independence between $(R_1, \ldots, R_p)$ and $(Y_1, \ldots, Y_p)$ assumed in $\mathcal{MS}$ is crucial to the result.

Instead of condition (4.3), let us now consider equality almost everywhere, i.e. pointwise equality except for a set of measure zero. We could assume the stronger condition

$$t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) \stackrel{a.e.}{=} t(Z_1, \ldots, Z_p),$$

for all $\alpha_1, \ldots, \alpha_p > 0$ and for all $Z = (Z_1, \ldots, Z_p)$ such that $Z = (R_1 Y_1, \ldots, R_p Y_p)$, where $R_i > 0$, Y_i takes values in $\mathcal{Y}_i$, and the marginal distribution of $(Y_1, \ldots, Y_p)$ is fixed. The latter can be shown to be equivalent to

$$t(Z_1, \ldots, Z_p) \stackrel{a.e.}{=} t(Y_1, \ldots, Y_p)$$

for all such Z, which implies that $t(Z)$ is distribution-free as Z ranges in this class. Observe that this result now holds in a wider class than $\mathcal{MS}$ since independence between $(R_1, \ldots, R_p)$ and $(Y_1, \ldots, Y_p)$ is no longer required.

As an illustration that the condition stated above is really stronger than (4.3), let us consider the following example for $p = 1$, where $\mathcal{MS}$ reduces to $\mathcal{S}$, and $n = 2$. We represent $z \in \Re^2$ through the usual polar coordinates $(w, r) \in [0, 2\pi) \times (0, \infty)$, and define $\mathcal{S}$ to be the spherical class, i.e. $Y = Z/\|Z\|_2$ is uniformly distributed on S^1. Consider the function

$$t(z) = \begin{cases} 1 & \text{if } (w, r) \in \{[0, \pi/2) \times (0, 1]\} \cup \{[\pi/2, \pi) \times (1, \infty)\} \\ 0 & \text{elsewhere.} \end{cases}$$

It is easily seen that (4.4) and thus (4.3) hold. However, if we consider the random variable in $\mathcal{S}$ corresponding to a Dirac distribution on 1 for R, and the set $A = \{z \in \Re^2 : w \in (0, \pi/2) \text{ and } r = 1\}$ with $P(A) = 1/4$, we obtain

$$t(\alpha z) \neq t(z), \text{ for all } z \in A, \alpha > 1,$$

and thus equality almost everywhere does not hold.

In practice, we shall often encounter functions $t(z)$ that only depend on $(y_1, \ldots, y_p)$ or, equivalently, the polar angles, and thus the distribution of $t(Z)$ can obviously only depend on the distribution of $(Y_1, \ldots, Y_p)$. Such functions of Z are thus trivially distribution-free in the wider class, where the marginal distribution of $(Y_1, \ldots, Y_p)$ is fixed without imposing independence of $(R_1, \ldots, R_p)$. In particular, Subsection 4.2 will provide some examples.

Let us now review some special cases of Theorem 1 that have appeared in the literature. If $\mathcal{Y}_1 = \ldots = \mathcal{Y}_p = S^{n-1}$, the unit sphere, and $Y_1, \ldots, Y_p$ are independently and uniformly distributed over S^{n-1}, Theorem 1 specializes to Theorem 5.1.1 (b) in Fang and Zhang (1990) for the multivariate spherical class.

In the case of one vector observation ($p = 1$), Theorem 1 reduces to Theorem 7.3 of Fang *et al.* (1990). Formally, their Theorem 7.3 is derived for random variables Z that have a stochastic representation $Z \stackrel{d}{=} RY$, where R is a positive random variable independent of Y, which 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. Of course, we could trivially extend Theorem 1 to this more general framework, as the proof does not rely on this representation of $\Re^n$. Our reason for adopting the representation in (2.1) with this one-to-one correspondence is to convey an interpretation to the random

variables R and Y. Such an interpretation seems instrumental in the use of our framework for practical modelling purposes, as was illustrated through the classes discussed in Section 3.

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

Proposition 1. *For the class $\mathcal{MS}$ defined in (4.1), all following statements are equivalent:*
(i) $t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p)$, for all $Z \in \mathcal{MS}$,
(ii) $t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p)$, for all $\alpha_1, \ldots, \alpha_p > 0$, for all $Z \in \mathcal{MS}$,
(iii) $t(\alpha_1 Y_1, \ldots, \alpha_p Y_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p)$, for all $\alpha_1, \ldots, \alpha_p > 0$,
(iv) $t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p)$, for all $Z \in \mathcal{MS}$, with $R_1, \ldots, R_p$ all independent,
(v) $t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p)$, for all $\alpha_1, \ldots, \alpha_p > 0$, for all $Z \in \mathcal{MS}$ with $R_1, \ldots, R_p$ all independent,
(vi) $t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p)$, for all $Z \in \mathcal{MS}$ with $R_1, \ldots, R_p$ all independent and such that $R_i \stackrel{d}{=} \sigma_i R_0$ for some positive random variable R_0 and some scalar $\sigma_i > 0$, $i = 1, \ldots, p$,
(vii) $t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p)$, for all $\alpha_1, \ldots, \alpha_p > 0$, for all $Z \in \mathcal{MS}$ with $R_1, \ldots, R_p$ all independent such that $R_i \stackrel{d}{=} \sigma_i R_0$ for some positive random variable R_0 and some scalar $\sigma_i > 0$, $i = 1, \ldots, p$,
(viii) $t(\alpha_1 Z_1, \ldots, \alpha_p Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p)$, for all $\alpha_1, \ldots, \alpha_p > 0$, for all $Z \in \mathcal{MS}$ with $R_1, \ldots, R_p$ independent and identically distributed.
In addition, any of (i)-(viii) implies
(ix) $t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(Y_1, \ldots, Y_p)$, for all $Z \in \mathcal{MS}$ with $R_1, \ldots, R_p$ independent and identically distributed,
which, in turn, implies
(x) $t(\alpha Z_1, \ldots, \alpha Z_p) \stackrel{d}{=} t(Z_1, \ldots, Z_p)$, for all $\alpha > 0$, for all $Z \in \mathcal{MS}$ with $R_1, \ldots, R_p$ independent and identically distributed.

Proof: Conditions (i)-(iii) are exactly (4.2)$-$(4.4), which have been shown to be equivalent in the context of Theorem 1. The other equivalences and implications are straightforward. •

In the case $p = 1$, Proposition 1 reduces to equivalence of Conditions (i), (ii) and (iii) and all other conditions become redundant. However, when $p > 1$, none of the ten conditions is redundant.

Conditions (iv) and (v) will be useful for the case of independent sampling. Note that $\mathcal{MS}$ will never exactly correspond to the class of $Z_1, \ldots, Z_p$ independently sampled from $\mathcal{S}$ in (3.1), since it does not preclude dependence among the R_i's. In order to have this interpretation of independent sampling from $\mathcal{S}$, we need to choose $\mathcal{MS}$ such that $Y_1, \ldots, Y_p$ are i.i.d., and we need to consider the subclass of this $\mathcal{MS}$ where $R_1, \ldots, R_p$ are all independent. Thus, Condition (v) provides us with a characterization of all distribution-free functions under independent sampling from a given class $\mathcal{S}$.

A much more interesting case from the practical perspective will consist in sampling independently from the same distribution but with possibly different scales. Under i.i.d. distributed $Y_1, \ldots, Y_p$, the characterization of distribution-freeness with this type of sampling from the corresponding $\mathcal{S}$ is analyzed through Conditions (vi) and (vii).

If we go to another practically relevant case, namely i.i.d. sampling, we focus on Conditions (viii)-(x). Again, choosing $\mathcal{MS}$ such that $Y_1, \ldots, Y_p$ are i.i.d., (viii) provides a sufficient condition and (x) gives a necessary condition for $t(Z_1, \ldots, Z_p)$ to be distribution-free under i.i.d. sampling from some random variable in $\mathcal{S}$. Since (viii) and (x) are not equivalent, we do not have a characterization of all distribution-free statistics in this case. Unfortunately, many practically useful statistics, like Hotelling's T^2 statistic, will verify (x) and thus distribution-freeness can not be excluded, but will not satisfy (viii), which implies that distribution-freeness can not be guaranteed. Thus, we have to remain inconclusive about the robustness properties of such statistics under i.i.d. sampling from $\mathcal{S}$. However, due to the equivalence between (viii) and (vi), we can conclude that they are not distribution-free under i.i.d. sampling up to a scale factor, from $\mathcal{S}$.

Thus, Proposition 1 provides results for three different schemes of repeated sampling from $\mathcal{S}$, namely, independent sampling, i.i.d. sampling up to a scale factor and pure i.i.d. sampling. 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, the first type of sampling would allow for each Z_i to be independently drawn from a different spherical distribution, e.g. Normal, Cauchy, Pearson type II, etc.; the second one 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 the third one really implies i.i.d. sampling from a particular spherical distribution, such as the standard Normal. Clearly, the last two cases seem the most interesting from a practical perspective.

In the coming Subsection we shall apply these findings to the context of inference robustness for a regression parameter. Moreover, we shall focus explicitly on the cases of i.i.d. error vectors and i.i.d. error vectors up to a scale factor, as they seem the most relevant for practical statistical applications. The reader will note that Proposition 1 directly allows for extending our results to alternative sampling schemes.

4.2. CLASSICAL INFERENCE ON REGRESSION PARAMETERS

In this Subsection we shall investigate robustness of sampling theory inference in the model

$$X_i = g_i(\beta) + \sigma_i \varepsilon_i, \quad i = 1, \ldots, p, \tag{4.5}$$

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

Important examples of (4.5) are the standard location-scale model, where $g_i(\beta) = \beta$

with $\beta \in \Re^n$, the case of a common location, where $g_i(\beta) = \beta\iota$ with scalar β and ι an n-dimensional vector of ones, and the regression context, where g_i 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. Subsection 5.1 will present a parallel Bayesian analysis of this model. We will examine two different situations:

(i) The case of equal scales, i.e. $\sigma_1 = \ldots = \sigma_p = \sigma$, which corresponds to $Z_1, \ldots, Z_p$ i.i.d.

(ii) The case of possibly different scales, which corresponds to $Z_1, \ldots, Z_p$ i.i.d. up to a scale factor.

Under the frequentist paradigm, inference on the common location parameter β in the model introduced in (4.5), will usually be based on the distribution of some function $t(\cdot)$ of $(Z_1, \ldots, Z_p) = (X_1 - g_1(\beta), \ldots, X_p - g_p(\beta))$. 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 would not depend on $(\sigma_1, \ldots, \sigma_p)$ either, $t(Z_1, \ldots, Z_p)$ would be 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.

In the case of equal scales in (4.5), i.e. $\sigma_1 = \ldots = \sigma_p = \sigma$, the random quantities $Z_1, \ldots, Z_p$ are themselves i.i.d. and Conditions (viii)-(x) of Proposition 1 immediately translate into the following sufficient and necessary conditions for distribution-freeness of $t(Z_1, \ldots, Z_p)$, expressed in terms of σ and $\varepsilon_1, \ldots, \varepsilon_p$.

Corollary 1. *Let $X_i = g_i(\beta) + \sigma\varepsilon_i$, $i = 1, \ldots, p$, where ε_i are i.i.d. random vectors with $\varepsilon_i \stackrel{d}{=} \varepsilon$ for some $\varepsilon \in \mathcal{S}$ in (3.1) , and let $t(\cdot)$ be a measurable function from $\Re^{n\times p}$ to $\Re^k$. If*

$$t(\alpha_1\{X_1 - g_1(\beta)\}, \ldots, \alpha_p\{X_p - g_p(\beta)\}) \stackrel{d}{=} t(X_1 - g_1(\beta), \ldots, X_p - g_p(\beta)),$$
for all $\alpha_1, \ldots, \alpha_p > 0$ and for all $\varepsilon \in \mathcal{S}$,

then

$$t(X_1 - g_1(\beta), \ldots, X_p - g_p(\beta)) \text{ has the same distribution}$$
for all $\varepsilon \in \mathcal{S}$,

which, in turn, implies

$$t(\alpha\{X_1 - g_1(\beta)\}, \ldots, \alpha\{X_p - g_p(\beta)\}) \stackrel{d}{=} t(X_1 - g_1(\beta), \ldots, X_p - g_p(\beta)),$$
for all $\alpha > 0$ and for all $\varepsilon \in \mathcal{S}$. •

Observe that, for $(Z_1, \ldots, Z_p) = (X_1 - g_1(\beta), \ldots, X_p - g_p(\beta))$, $t(Z_1, \ldots, Z_p)$ having the same distribution for all $\varepsilon \in \mathcal{S}$, immediately leads to the following two consequences of practical interest:

First, the distribution of $t(Z_1, \ldots, Z_p)$ does not depend on the scale parameter σ, and $t(Z_1, \ldots, Z_p)$ is thus a pivotal quantity, that can be used for inference on β when σ is a nuisance parameter.

In addition, the distribution of $t(Z_1, \ldots, Z_p)$ is exactly the same for all possible choices of ε in the class $\mathcal{S}$. As an example, if $\mathcal{S}$ is the spherical class, it would not matter whether ε has a Normal distribution or any other spherical distribution. We would thus achieve robustness with respect to deviations from Normality in the spherical class.

Corollary 1 tells us that the sufficient condition for distribution-freeness 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 $\varepsilon \in \mathcal{S}$. Obviously, the necessary and sufficient conditions in Corollary 1 coincide when $p = 1$, i.e. in the case of one vector observation. However, for $p > 1$ we do not have a characterization of all distribution-free pivotal quantities $t(Z_1, \ldots, Z_p)$, and the mere fact that $t(Z_1, \ldots, Z_p)$ is a pivot does not guarantee distribution-freeness. The crucial assumption that prevents us from obtaining a full characterization is that the Z_i's share exactly the same distribution, i.e. not only the same distributional shape, but also the same scale. If we relax the latter requirement by allowing for possibly different scales σ_i, while retaining the former, we can easily deduce from Conditions (vi)-(vii) in Proposition 1:

Corollary 2. *Let $X_i = g_i(\beta) + \sigma_i \varepsilon_i$, $i = 1, \ldots, p$, where ε_i are i.i.d. random vectors with $\varepsilon_i \stackrel{d}{=} \varepsilon$ for some $\varepsilon \in \mathcal{S}$ in (3.1) , and let $t(\cdot)$ be a measurable function from $\Re^{n \times p}$ to $\Re^k$. Then*

$$t(X_1 - g_1(\beta), \ldots, X_p - g_p(\beta)) \text{ has the same distribution}$$
$$\text{for all } \sigma_1, \ldots, \sigma_p > 0 \text{ and for all } \varepsilon \in \mathcal{S},$$

if and only if

$$t(\alpha_1\{X_1 - g_1(\beta)\}, \ldots, \alpha_p\{X_p - g_p(\beta)\}) \stackrel{d}{=} t(X_1 - g_1(\beta), \ldots, X_p - g_p(\beta)),$$
$$\text{for all } \alpha_1, \ldots, \alpha_p > 0 \text{ and for all } \varepsilon \in \mathcal{S}. \quad \bullet$$

Observe that, for $(Z_1, \ldots, Z_p) = (X_1 - g_1(\beta), \ldots, X_p - g_p(\beta))$, distribution-freeness of $t(Z_1, \ldots, Z_p)$ for $\sigma_1, \ldots, \sigma_p > 0$ implies that this function is a pivotal quantity that can be used for inference on β in the presence of nuisance scale parameters. In addition, distribution-freeness for $\varepsilon \in \mathcal{S}$ implies 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 $\mathcal{S}$.

Corollary 2 tells us that $t(Z_1, \ldots, Z_p)$ is a distribution-free pivot for all $\varepsilon \in \mathcal{S}$ if and only if the distribution of $t(Z_1, \ldots, Z_p)$ is invariant with respect to arbitrary changes of scale of each of the columns of $(Z_1, \ldots, Z_p)$ for any $\varepsilon \in \mathcal{S}$. For the model with possibly different scales σ_i, $i = 1, \ldots, p$, this latter property is equivalent to saying that, for any $\varepsilon \in \mathcal{S}$, the distribution of $t(Z_1, \ldots, Z_p)$ does not depend on $(\sigma_1, \ldots, \sigma_p)$, i.e. $t(Z_1, \ldots, Z_p)$ is a pivotal quantity. Therefore, Corollary 2 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 $\varepsilon \in \mathcal{S}$.

Due to the equivalence of Conditions (i)-(vii) in Proposition 1, we know that the characterization in Corollary 2 automatically extends to the much wider classes where $Z_1, \ldots, Z_p$ are drawn independently from $\mathcal{S}$, or even $(Z_1, \ldots, Z_p) \in \mathcal{MS}$, where $\mathcal{MS}$ is the

matricvariate class in (4.1) corresponding to $\mathcal{S}$. However, in statistical practice one will usually be interested in the situation of independent replications of the same experiment, up to, at most, scale changes, and thus we will not consider these wider classes.

Trivially, sets of distribution-free functions of $(Z_1, \ldots, Z_p)$ are never empty, because all constant functions belong to them. However, we obviously look for nontrivial elements of these sets. Often we possess useful pivotal quantities for the case of just 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 a certain set $\mathcal{S}$ of the type introduced in (3.1), 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

$$t(z_1, \ldots, z_p) = f\{t_1(z_1), \ldots, t_p(z_p)\}, \text{ for } (z_1, \ldots, z_p) \in \Re^{n \times p}.$$

Then we obtain

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

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-free 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 some class $\mathcal{S}$. This implies that, for any measurable $f(\cdot)$, $f\{t_1(Z_1), \ldots, t_p(Z_p)\}$ is distribution-free not only for $Z_1, \ldots, Z_p$ independently sampled from $\mathcal{S}$, but from the much larger class of multivariate distributions

$$\mathcal{C} = \{Z : Z = RY \text{ through (2.1)}, \text{ with } Y \text{ distributed as } P_2\}, \tag{4.6}$$

characterized by the same marginal probability distribution P_2 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 4.1: Linear regression under independent sampling.

Consider the sampling model for X_i, random vectors in $\Re^n$,

$$X_i = D_i\beta + \sigma_i\epsilon_i, \quad i = 1, \ldots, p,$$

which is a special case of (4.5) with $g_i(\beta) = D_i\beta$, where D_i is an $n \times m$ fixed matrix, $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d. with $\varepsilon_i \stackrel{d}{=} \varepsilon$, a random vector with a standard n-variate Normal

distribution, $\sigma_i > 0$ is a (possibly observation-specific) scale parameter, and β 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 full column rank for D_i and $n > m$,

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

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, functions of $(S_1, \ldots, S_p)$ are pivotal quantities, 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

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

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-free whenever ϵ is in a given class $\mathcal{C}$ of n-variate random variables characterized by some marginal distribution of the polar angles, or, equivalently, a probability distribution on some space $\mathcal{Y}$ as described in Section 2. Of course, this marginal distribution on $\mathcal{Y}$ should be chosen such that $P\{\epsilon \in A_i\} = 0$ for all $i = 1 \ldots, p$, and for all $\varepsilon \in \mathcal{C}$. Due to the fact that the scale invariance property of $t_i(\cdot)$ holds everywhere, and not only in distribution, we obtain robustness in the entire class $\mathcal{C}$ in (4.6). 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 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 making classical inferences on β. It should be clear, however, that robustness of inferences based on $(S_1, \ldots, S_p)$ is achieved at the cost of efficiency losses in the case of equal scale factors, where we know that $\sigma_1 = \ldots = \sigma_p = \sigma$. The latter information is not used by $(S_1, \ldots, S_p)$ which is

constructed as if the scales were different, and is thus invariant to individual rescaling of each error vector $X_i - D_i\beta$.

In the case of a single Normal vector observation, say X_i, some functions of $S_i(\beta)$ are not just arbitrary pivots, but follow from some general principles of classical inference. For instance, it is well-known that the statistic $F_i(\beta_0) \equiv S_i(\beta_0)'D_i'D_iS_i(\beta_0)/m$ can be derived using the likelihood ratio principle for testing $H_0 : \beta = \beta_0$ against $H_1 : \beta \in \Re^m - \{\beta_0\}$. Therefore, this test statistic has a particular interpretation in the reference case of Normality, and it keeps its $F(m, n-m)$ distribution under the null hypothesis whenever ε_i has a non-Normal spherical distribution.

The likelihood ratio principle, when applied to some reference distribution, often leads to robust test statistics and pivotal quantities in the case of one vector observation. 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.

Example 4.2: Distribution-free pivots for 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 β is the scalar parameter of interest and $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d. with $\varepsilon_i \stackrel{d}{=} \varepsilon$, an n-dimensional random vector. In this example we will examine robustness when ε follows an l_q-spherical distribution, introduced in Subsection 3.1.2, for a fixed value of $q \in (0, \infty)$. The case $q = \infty$ can be analyzed in a similar fashion.

We first consider just one vector observation, $X_i = (X_{1i}, \ldots, X_{ni})'$. In the reference case where the elements of ε_i are i.i.d. draws from an exponential power distribution, the likelihood function is given by

$$l(\beta, \sigma_i; X_i) \propto \sigma_i^{-n} \exp\left\{ -\frac{1}{2}\sigma_i^{-q} \sum_{j=1}^{n} |X_{ji} - \beta|^q \right\}.$$

The likelihood ratio principle for testing $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ leads to the statistic

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

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

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

and, in particular, is the median of $X_{1i}, \ldots, X_{ni}$ if $q = 1$, and the sample mean, $\overline{X}_i$, if $q = 2$. It is easy to check that $LR(\beta_0; X_i)$ is a function of $X_i - \beta_0\iota$, say $t(X_i - \beta_0\iota)$, and that $t(\alpha z) = t(z)$, for all $\alpha > 0$ and all $z \in \Re^n - \{\iota\gamma : \gamma \in \Re\}$.

Thus, the likelihood ratio principle leads to the scale invariant pivotal quantity $t(X_i - \beta\iota) = LR(\beta; X_i)$, which is therefore distribution-free 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. In particular, $LR(\beta; X_i)$ will have the same distribution for any ε_i in the l_q-spherical subclass $\mathcal{S}$ of $\mathcal{C}$. Any measurable function $f\{LR(\beta; X_1), \ldots, LR(\beta; X_p)\}$ will now define a pivotal quantity which will be distribution-free when ε ranges in the l_q-spherical class, and thus potentially useful for robust inference on β. The choice of $f(\cdot)$ remains an open problem.

Of course, except for $q = 2$, the sampling distribution of $LR(\beta; X_i)$ and of functions thereof has to be investigated through numerical techniques.

4.3. DISTRIBUTION AND INFERENCE INVARIANCE IN $\mathcal{R}$

In this Subsection, we shall analyze the extent to which exact robustness results can be characterized under repeated sampling from the class $\mathcal{R}$ defined in (3.12).

Since inference will often be based on repeated sampling from vector observations, it is natural to consider matrix random variables. In particular, we shall again focus on matrices $Z = (Z_1, \ldots, Z_p)$, where each column Z_i is a random variable in $\Re^n - \{0\}$ represented as $Z_i = R_i Y_i$ through (2.1), with R_i taking values in $\Re_+$ and Y_i in a space $\mathcal{Y}_i$. We then define

$$\mathcal{MR} = \{Z = (Z_1, \ldots, Z_p) : Z = (R_1 Y_1, \ldots, R_p Y_p) \text{ with } (R_1, \ldots, R_p) \text{ independent of } (Y_1, \ldots, Y_p) \text{ and the distribution of } (R_1, \ldots, R_p) \text{ fixed}\}. \tag{4.7}$$

Obviously, for $p = 1$, $\mathcal{MR}$ reduces to the class $\mathcal{R}$ in (3.12). Also note that if $\mathcal{Y}_i = \mathcal{Y}$, $i = 1, \ldots, p$, and the fixed joint distribution of $(R_1, \ldots, R_p)$ leads to the same marginal distribution for each R_i, then $Z_1, \ldots, Z_p$ are all in the same class $\mathcal{R}$.

Alternatively, we could use the representation in (2.2) and characterize Z through $Z = (R_1 k_1(W_1), \ldots, R_p k_p(W_p))$, where each k_i $(i = 1, \ldots, p)$ is a one-to-one transformation from $\mathcal{W} \subset \Re^{n-1}$ to $\mathcal{Y}_i$. Therefore, we could equivalently define $\mathcal{MR}$ in terms of the latter representation by fixing the distribution of $(R_1, \ldots, R_p)$ and imposing independence between $(R_1, \ldots, R_p)$ and $(W_1, \ldots, W_p)$. In fact, this equivalent representation will prove to be more convenient for the subsequent discussion, where we shall assume, without loss of generality, that the $(n-1)$-dimensional vector of zeros belongs to $\mathcal{W}$.

We are now concerned with characterizing the measurable functions $t(\cdot)$, taking values in $\Re^k$, that are distribution-free in $\mathcal{MR}$, or in certain subsets of $\mathcal{MR}$. The following Theorem provides some useful results in this respect.

Theorem 2. *Consider the class $\mathcal{MR}$ in (4.7) and some measurable function $t(\cdot)$ taking values in $\Re^k$. The following statements are equivalent:*

(i) $t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(R_1 k_1(0), \ldots, R_p k_p(0))$, for all $Z \in \mathcal{MR}$,

(ii) $t(R_1 k_1(w_1), \ldots, R_p k_p(w_p)) \stackrel{d}{=} t(R_1 k_1(0), \ldots, R_p k_p(0))$, for all $w_1, \ldots, w_p \in \mathcal{W}$,

(iii) $t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(R_1 k_1(0), \ldots, R_p k_p(0))$, for all $Z \in \mathcal{MR}$, with $W_1, \ldots, W_p$ all independent.

In addition, any of (i)-(iii) implies

(iv) $t(Z_1, \ldots, Z_p) \stackrel{d}{=} t(R_1 k_1(0), \ldots, R_p k_p(0))$, for all $Z \in \mathcal{MR}$, with $W_1, \ldots, W_p$ independent and identically distributed,

which, in turn, implies

(v) $t(R_1 k_1(w), \ldots, R_p k_p(w)) \stackrel{d}{=} t(R_1 k_1(0), \ldots, R_p k_p(0))$, *for all* $w \in \mathcal{W}$.

Proof: see Appendix. •

As was the case in Theorem 1, the proof of Theorem 2 does not rely on the one-to-one correspondence between points $z \in \Re^n - \{0\}$ and pairs (y, r), established in Section 2. Therefore, the Theorem could alternatively be stated and proved without this restriction.

Condition (i) means that $t(Z)$ is distribution-free in the entire class $\mathcal{MR}$, whereas Condition (ii) indicates distribution-freeness in the subclass of $\mathcal{MR}$ corresponding to all Dirac distributions for $(W_1, \ldots, W_p)$. Thus, $t(Z)$ is distribution-free in all of $\mathcal{MR}$ if and only if it has that property in this much smaller subclass. This, again, derives from the fact that under independence between $(R_1, \ldots, R_p)$ and $(W_1, \ldots, W_p)$, $\mathcal{MR}$ is the convex hull of its subclass generated by all Dirac distributions for $(W_1, \ldots, W_p)$.

If we focus on independent sampling, we naturally consider (iii). Choosing $k_i(\cdot) = k(\cdot)$ $(i = 1, \ldots, p)$ and $\mathcal{MR}$ such that $R_1, \ldots, R_p$ are i.i.d., Condition (iii) states that $t(Z_1, \ldots, Z_p)$ is distribution-free under independent (not necessarily identical) sampling of $Z_1, \ldots, Z_p$ from the corresponding class $\mathcal{R}$, defined in (3.12). Clearly, this will imply that $t(Z_1, \ldots, Z_p)$ is distribution-free if we additionally restrict ourselves to sampling independently from the same distribution in $\mathcal{R}$ (i.i.d.), as is stated in Condition (iv). Thus, either of (i)-(iii) provide sufficient conditions for distribution-freeness under i.i.d. sampling from $\mathcal{R}$, whereas Condition (v) is merely necessary for such robustness. Again, like in Proposition 1, we do not obtain a characterization of robustness in the case of pure i.i.d. sampling.

In contrast to Theorem 1, Theorem 2 does not provide us with a characterization of distribution-freeness in terms of invariance under transformations of the observables. In Subsection 4.1, such invariance was described in (4.3). In order to obtain a similar condition for the class $\mathcal{MR}$, we need to find a set of transformations, denoted by $\mathcal{Q}$, such that the condition

$$t(Q(Z)) \stackrel{d}{=} t(Z), \text{ for all } Z \in \mathcal{MR}, \text{ for all } Q \in \mathcal{Q} \tag{4.8}$$

holds if and only if (i), (ii) and (iii) apply. Then (4.8) would be a characterization of distribution-freeness in $\mathcal{MR}$ in terms of transformations Q on the observables. It can be shown that one choice for $\mathcal{Q}$ satisfying the above would be the set of all possible transformations that only affect the angles of each column of $Z \in \mathcal{MR}$, i.e.

$$Q(Z) = Q(R_1 k_1(W_1), \ldots, R_p k_p(W_p)) \equiv (R_1 k_1(W_1^*), \ldots, R_p k_p(W_p^*)), \tag{4.9}$$

where $W_1^*, \ldots, W_p^*$ are any random variables taking values in $\mathcal{W}$. However, the class described through (4.9) is so large that it does not lead to a characterization that is easy to check in practice. Similarly, for robustness in $\mathcal{MS}$ we could have considered in Subsection 4.1, instead of the transformations in (4.3), all possible transformations that only affect the radius of each column of Z. However, (4.3) is still equivalent to (4.2) and is very easy to verify in practice and to interpret. Thus, we would ideally want to characterize robustness

in $\mathcal{MR}$ through a subclass of $\mathcal{Q}$ defined through (4.9), which still preserves equivalence with (i), (ii) and (iii) in Theorem 2, yet makes (4.8) easy to check if we restrict it to this subclass.

In the case where each $k_i(\cdot) = h(\cdot)$, $i = 1, \ldots, p$, where $\{h(w) : w \in \mathcal{W}\} = S^{n-1}$, the unit sphere, such as in repeated sampling from the anisotropic class, a natural subclass with these properties is given by all the transformations Q such that

$$Q(Z) = (\Gamma_1 Z_1, \ldots, \Gamma_p Z_p),$$

where $\Gamma_1, \ldots, \Gamma_p \in \mathcal{O}(n)$ are any orthogonal matrices. If we consider the general case, a possible choice would be the subclass of (4.9) characterized by $W_i^* = h^{-1}\{\Gamma_i h(W_i)\}$, $i = 1, \ldots, p$ and $\Gamma_i \in \mathcal{O}(n)$. This corresponds to first transforming from $\mathcal{Y}_i = \{k_i(w) : w \in \mathcal{W}\}$ to S^{n-1}, then multiplying by any orthogonal matrix, and finally transforming back to $\mathcal{Y}_i$, for each $i = 1, \ldots, p$.

In some applied statistics problems, pivotal quantities can be found that are only a function of $(R_1, \ldots, R_p)$. Clearly, such quantities will be distribution-free as long as the marginal distribution of $(R_1, \ldots, R_p)$ is fixed. Thus, we will directly obtain invariance in $\mathcal{MR}$, and also in the wider class where independence between $(R_1, \ldots, R_p)$ and $(Y_1, \ldots, Y_p)$ is no longer imposed. A context in which such a situation naturally appears is that of i.i.d. sampling from the scale model

$$X_i = \sigma \varepsilon_i, \quad i = 1, \ldots, p, \tag{4.10}$$

where $\sigma > 0$ is the scale parameter, as illustrated by the following example.

Example 4.3: Independent sampling from an l_q-anisotropic scale model.

We assume that the observations are generated from the scale model in (4.10), where each n-vector ε_i is the result of independent sampling from an exponential power distribution, with a fixed $q \in (0, \infty]$. Furthermore, all vectors $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d.

The distributional assumption on each ε_i implies that the q^{th} power of its l_q-radius $v_q(\varepsilon_i)$, defined in Subsection 3.1.2, has a $\chi^2_{2n/q}$ distribution for finite q, and the l_∞-radius is Beta$(n, 1)$ distributed. Thus, for $0 < q < \infty$,

$$t_i\left(\frac{X_i}{\sigma}\right) \equiv \left(\frac{v_q(X_i)}{\sigma}\right)^q \sim \chi^2_{2n/q}, \quad i = 1, \ldots, p,$$

and taking the sum over all the observations

$$t\left(\frac{X_1}{\sigma}, \ldots, \frac{X_p}{\sigma}\right) \equiv \sum_{i=1}^{p} t_i\left(\frac{X_i}{\sigma}\right) = \sigma^{-q} \sum_{i=1}^{p} \{v_q(X_i)\}^q \sim \chi^2_{2np/q}.$$

In the case of $q = \infty$, we know

$$t_i^*\left(\frac{X_i}{\sigma}\right) \equiv \frac{\max_{j=1,\ldots,n} |X_{ji}|}{\sigma} \sim \text{Beta}(n, 1), \quad i = 1, \ldots, p,$$

where $X_i = (X_{1i}, \ldots, X_{ni})$ for $i = 1, \ldots, p$, leading to

$$t^*\left(\frac{X_1}{\sigma}, \ldots, \frac{X_p}{\sigma}\right) \equiv \max_{i=1,\ldots,p} t_i^*\left(\frac{X_i}{\sigma}\right) = \frac{\max_{j=1,\ldots,n;i=1,\ldots,p} |X_{ji}|}{\sigma} \sim \text{Beta}(np, 1).$$

The pivotal quantities $t(X_1/\sigma, \ldots, X_p/\sigma)$ and $t^*(X_1/\sigma, \ldots, X_p/\sigma)$ can now be used for inference on σ.

Writing $\varepsilon_i = R_i Y_i$ with $R_i = v_q(\varepsilon_i)$ and $Y_i = \varepsilon_i / v_q(\varepsilon_i)$ as in Subsection 3.1.2, we note that $t(\cdot)$ and $t^*(\cdot)$ only depend on ε_i throught its l_q-radius R_i. Thus, the distribution of these pivotal quantities will be the same as long as the distribution of R_i is fixed. If, in addition, we impose independence between R_i and Y_i, we obtain perfect robustness whenever ε_i ranges in the l_q-anisotropic class, introduced in Subsection 3.2, for given q.

5. ROBUSTNESS OF BAYESIAN INFERENCE

In this Section we shall consider the issue of robust inference from a Bayesian angle. The nature of the Bayesian paradigm will lead to a somewhat different approach, not based in the distribution theory of Subsection 4.1, but rather in the joint distribution of observables and parameters. We shall consider both the regression model in (4.5) (Subsection 5.1) and the scale model in (4.10) (Subsection 5.2).

5.1. THE REGRESSION MODEL

5.1.1. Inference on location and regression parameters

Let us consider again the model in (4.5), $X_i = g_i(\beta) + \sigma_i \varepsilon_i$, $i = 1, \ldots, p$, where $\varepsilon_i, \ldots, \varepsilon_p$ are i.i.d. n-dimensional random vectors (such that the conditional distribution of ε_i given (β, σ_i) does not depend on these parameters), $\sigma_i > 0$, $i = 1, \ldots, p$, are scale parameters and $g_i(\beta)$, $i = 1, \ldots, p$, are location vectors, parameterized in terms of a common vector $\beta \in \mathcal{B} \subset \Re^m$ $(m \leq n)$ through known functions $g_i(\cdot)$ from $\Re^m$ to $\Re^n$.

In Corollaries 1 and 2 of Subsection 4.2 we examined robustness of classical inference procedures when ε_i is a random variable from the class $\mathcal{S}$ in (3.1). As will be clear in the sequel, our Bayesian robustness results will be obtained for the wider class $\mathcal{C}$, defined in (4.6), where independence between R and Y is not required. Therefore, we shall now take $\varepsilon_1, \ldots, \varepsilon_p$ to be i.i.d. random vectors from $\mathcal{C}$.

As in Subsection 4.2, we will consider both the case of equal scale factors, i.e. $\sigma_1 = \ldots = \sigma_p = \sigma$, and the case of possibly different scales. In both situations we will assume Jeffreys' improper prior density on the scale parameter(s), which conveys the idea of prior ignorance about scale. The following Theorem constitutes the basis for deriving Bayesian robustness results.

Theorem 3. *Consider the general multivariate location-scale model $X = \mu + \sigma\varepsilon$, where ε is an n-dimensional random vector represented as $\varepsilon = RY$ through (2.1), $\sigma > 0$ is a scale parameter and $\mu \in \mathcal{M} \subset \Re^n$ is a location parameter.*

Let us choose a prior distribution for (μ, σ), which is the product measure corresponding to the improper density $p(\sigma) = c\sigma^{-1}$ $(c > 0)$ for σ and any σ-finite prior measure for μ. Then the joint distribution of X and μ does not depend on the conditional probability distribution of R given Y.

Proof: see Appendix. •

From the proof it is easy to see that the key to this result is the invariance of the density $p(\sigma) = c\sigma^{-1}$ under the group of transformations $\{\sigma r : r > 0\}$. As a consequence of this, it is obtained that $(\mu, \sigma R, Y, R)$ has the same distribution as (μ, σ, Y, R), from which can be derived that the distribution of $(\mu, \sigma R, Y)$ does not depend on the conditional probability of R given Y. Thus, the distribution of $(\sigma RY, \mu) = (X - \mu, \mu)$, and therefore that of (X, μ), do not depend on the conditional probability of R given Y either. The improper density $p(\sigma) \propto \sigma^{-1}$ is the only one with this invariance property.

Observe that, paralleling the discussions following Theorems 1 and 2, the proof of this result does not make use of the unique representation of points $z \in \Re^n$ in terms of pairs (y, r). Therefore, Theorem 3 also holds in the more general situation where $\varepsilon \stackrel{d}{=} RY$ for some positive random variable R. In contrast to Theorems 1 and 2, independence between R and Y is now not required.

Theorem 3 implies that, under the Jeffreys' prior for the scale parameter σ, the joint distribution of (X, μ) only depends on the distribution of ε through the marginal probability of Y. Therefore, if we fix the distribution of Y and we consider the corresponding class $\mathcal{C}$ in (4.6), the joint distribution of (X, μ) will be exactly the same for any choice of $\varepsilon \in \mathcal{C}$. In addition, if the marginal distribution of X is σ-finite, the posterior probability distribution of μ given X is well-defined, and will also be unaffected by the particular choice of $\varepsilon \in \mathcal{C}$. Thus, posterior inference on μ, whenever it can be conducted, is perfectly robust when ε ranges in an entire class $\mathcal{C}$. This result generalizes the finding in Fernández *et al.* (1994), which treated the special case where all the distributions are dominated by the Lebesgue measure in the corresponding space.

So far, we have presented a Bayesian robustness result for the case of one single vector observation. We now examine the case of independent sampling. In the practically relevant situation of independent sampling from (4.5) with unknown and possibly different scale factors, perfect inference robustness is easily derived from Theorem 3. The result is stated in the following Corollary.

Corollary 3. *Consider the sampling model*

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

where $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d. n-dimensional random vectors represented as $\varepsilon_i = R_i Y_i$ through (2.1), $\sigma_i > 0$, $i = 1, \ldots, p$, are scale parameters and $g_i(\beta)$, $i = 1, \ldots, p$, are location vectors, parameterized in terms of a common vector $\beta \in \mathcal{B} \subset \Re^m (m \leq n)$ through known measurable functions $g_i(\cdot)$ 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 $(X_1, \ldots, X_p, \beta)$ *does not depend on the conditional distribution of* R_i *given* Y_i.

Proof: The proof of this Corollary parallels that of Theorem 3. We now obtain that $(\beta, \sigma_1 R_1, Y_1, R_1, \ldots, \sigma_p R_p, Y_p, R_p)$ has the same distribution as $(\beta, \sigma_1, Y_1, R_1, \ldots, \sigma_p, Y_p, R_p)$. As a consequence, the distribution of $(\beta, \sigma_1 R_1, Y_1, \ldots, \sigma_p R_p, Y_p)$, and thus the distribution of $(X_1, \ldots, X_p, \beta)$, do not depend on the conditional probability of R_i given Y_i. •

Corollary 3 implies that, under the standard non-informative prior on the scale parameters, we obtain exactly the same distribution of $(X_1, \ldots, X_p, \beta)$ for any choice of ε_i in a given class $\mathcal{C}$ of the type described in (4.6). Therefore posterior inference on β and predictive inference, provided they can be conducted, are perfectly robust with respect to deviations of ε_i within an entire class $\mathcal{C}$. This Bayesian robustness parallels the classical robustness result of Corollary 2, although it is now obtained for the wider class $\mathcal{C}$.

In the case that the class $\mathcal{C}$ considered corresponds to some density $f_2(w)$ on the polar angles $w \in \mathcal{W}$, 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

$$p(x_1, \ldots, x_p, \beta) = cp(\beta) \prod_{i=1}^{p} \frac{\|x_i - g_i(\beta)\|_2^{-n}}{s\left\{h^{-1}\left(\frac{x_i - g_i(\beta)}{\|x_i - g_i(\beta)\|_2}\right)\right\}} f_2\left\{h^{-1}\left(\frac{x_i - g_i(\beta)}{\|x_i - g_i(\beta)\|_2}\right)\right\}, \qquad (5.1)$$

where $\{h(w) : w \in \mathcal{W}\} = S^{n-1}$ and $u^{n-1}s(w)$ is the Jacobian of the polar transformation (u denotes the Euclidean radius).

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 show that in most practically relevant situations this requirement will not be met in the pure location-scale model, where $g_i(\beta) = \beta$, $i = 1, \ldots, p$, and $\beta \in \Re^n$ is not restricted to a lower dimensional subspace.

Proposition 2. *Consider* p *independent observations from the multivariate location-scale model*

$$X_i = \beta + \sigma_i \varepsilon_i, \quad i = 1, \ldots, p,$$

where $\varepsilon_1, \ldots, \varepsilon_p$ *are i.i.d.* n*-dimensional random vectors from a class* $\mathcal{C}$ *characterized by a density* $f_2(w)$ *on the polar angles* $w \in \mathcal{W}$.
We assume the prior density

$$p(\beta, \sigma_1, \ldots, \sigma_p) = p(\beta) \prod_{i=1}^{p} p(\sigma_i) = cp(\beta) \prod_{i=1}^{p} \sigma_i^{-1} \quad (c > 0), \qquad (5.2)$$

where $p(\beta)$ *is any density, either proper or* σ*-finite, that does not restrict* β *to a lower dimensional subspace of* $\Re^n$.
If there exists some positive constant K *such that* $f_2(w)/s(w) > K$ *for all* $w \in \mathcal{W}$, *then the predictive distribution of* $(X_1, \ldots, X_p)$ *is not* σ*-finite.*

Proof: see Appendix. •

As an immediate consequence of Proposition 2, we will typically not be able to conduct posterior inference on β in pure location-scale models under the prior structure in (5.2), which ensures perfect robustness of $p(x_1, \ldots, x_p, \beta)$ for $\varepsilon_i \in \mathcal{C}$. A sufficient condition preventing such inference is that $f_2(w)/s(w) > K > 0$ for all $w \in \mathcal{W}$ and some positive constant K. Alternatively, we can rewrite this sufficient condition in terms of the function $v(\cdot)$ associated with $f_2(\cdot)$ through (3.10) and (3.11), as

$$v\{h(w)\} < K' < \infty, \text{ for all } w \in \mathcal{W} \text{ and some positive constant } K'.$$

This covers most of the practically relevant cases, such as all classes $\mathcal{C}$ associated to e.g. the spherical class, any l_q-spherical class or any v-spherical class whenever $v(\cdot)$ is a norm or the isodensity contours are bounded away from the origin.

This clearly shows that the perfect robustness of $p(x_1, \ldots, x_p, \beta)$ with respect to the form of the distribution of ε_i in the wide class $\mathcal{C}$ is sometimes achieved in cases where inference on the location parameter is precluded.

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 5.1. Independent linear regressions.

Consider the same sampling model as in Example 4.1, i.e. p independent Normal regression equations $X_i = D_i\beta + \sigma_i\varepsilon_i$, and the prior structure as in (5.2).

The resulting Bayesian model was studied by Tiao and Zellner (1964), Dickey (1968), Zellner (1971), 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 4.1, $i = 1, \ldots, p$. 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 4.1, considered as a function of β. Such a posterior density is called a $p-0$ poly-t density [see Drèze (1977)], and for $m < n$ is clearly proper. 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 densities were obtained as the posteriors resulting from independent Normal samples with different variances. Corollary 3 shows that the assumption of Normality is not necessary. Under the prior structure in (5.2), 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.

In order to obtain the result in Corollary 3, it was crucial to assume p unrelated scale factors σ_i with Jeffreys' improper prior on each of them. As the density $p(\sigma_i) \propto \sigma_i^{-1}$ is invariant under the group of transformations $\{r_i\sigma_i : r_i > 0\}$, each scale factor σ_i absorbes the influence of the corresponding R_i. The latter is not possible in the case of equal scales, i.e.

$$X_i = g_i(\beta) + \sigma\varepsilon_i, \quad i = 1, \ldots, p,$$

where $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d. n-dimensional random vectors represented as $\varepsilon_i = R_iY_i$ through (2.1), $\sigma > 0$ is a common scale parameter and $\beta \in \mathcal{B} \subset \Re^m$ is a common location (or regression) parameter. In this case, it is impossible to achieve robustness when ε_i ranges in a class $\mathcal{C}$ of the type described in (4.6). To illustrate this fact, let us consider the special situation where all measures considered are dominated by Lebesgue measure in the corresponding space. The prior density will take the form

$$p(\beta, \sigma) = p(\beta)p(\sigma) = p(\beta)c\sigma^{-1}, \quad (c > 0).$$

We shall consider the density of ε_i, represented as $\varepsilon_i = R_ik(W_i)$ through (2.2), in terms of its coordinates (W_i, R_i), factorized as the product of the marginal density of W_i, $f_2(w_i)$, and the conditional density of R_i given W_i, $f_1(r_i; w_i)$. Thus, the joint density of $(\beta, \sigma, R_1, W_1, \ldots, R_p, W_p)$ is given by

$$p(\beta, \sigma, r_1, w_1, \ldots, r_p, w_p) = p(\beta)c\sigma^{-1} \prod_{i=1}^{p} f_1(r_i; w_i)f_2(w_i).$$

Transforming from $(\beta, \sigma, R_1, W_1, \ldots, R_p, W_p)$ to $(\beta, \sigma, \lambda_1, W_1, \ldots, \lambda_p, W_p)$, where $\lambda_i = \sigma R_i$ is the radial coordinate of $X_i - g_i(\beta)$, $i = 1, \ldots, p$, leads to

$$p(\beta, \sigma, \lambda_1, w_1, \ldots, \lambda_p, w_p) = cp(\beta)\frac{1}{\sigma^{p+1}} \prod_{i=1}^{p} f_1\left(\frac{\lambda_i}{\sigma}; w_i\right) f_2(w_i).$$

Therefore, $p(\beta, \lambda_1, w_1, \ldots, \lambda_p, w_p)$, and thus $p(x_1, \ldots, x_p, \beta)$, depend on the form of $f_1(\cdot)$ and we do not have robustness with respect to the choice of $\varepsilon_i \in \mathcal{C}$.

From this discussion we conclude that in the case of a common scale parameter, which corresponds to i.i.d. error vectors $Z_i = \sigma\varepsilon_i$, robustness of Bayesian inference in a class $\mathcal{C}$ in (4.6) seems impossible when $p > 1$. The crucial difference with the situation in Corollary 3 that prevents robustness in this case, is that now there is just one scale factor σ which can only absorb the influence of one of the R_i's.

5.1.2. Inference on scale parameters

In Subsection 5.1.1 we proved that for the model in (4.5), with the prior structure in (5.2), the distribution of $(X_1, \ldots, X_p, \beta)$ is exactly the same whenever ε_i ranges in a given class $\mathcal{C}$ as defined in (4.6). In this Subsection, we shall analyze inference on scale in the same model.

We start from the assumptions in Theorem 3, i.e. we consider the n-variate location-scale model, $X = \mu + \sigma\varepsilon$, where $\varepsilon = RY$ through (2.1), and a prior distribution on (μ, σ) which is the product measure corresponding to the improper density $p(\sigma) = c\sigma^{-1}$ $(c > 0)$ for σ and any σ-finite prior for μ. From the proof of Theorem 3 follows that the marginal distribution of (μ, λ, Y), where $\lambda = \sigma R$, derived from the joint σ-finite distribution of (R, μ, λ, Y), is σ-finite. This implies that the conditional probability distribution of R given (μ, λ, Y) is well-defined. Due to the one-to-one correspondence between (R, μ, λ, Y) and $(\sigma, \mu, \lambda, Y)$, we conclude that the conditional distribution of σ given (μ, λ, Y) is also well-defined. In particular, for any Borel measurable set $A \subset \Re_+$,

$$P_{\sigma|(\mu,\lambda,y)}\{A\} = P_{R|(\mu,\lambda,y)}\left\{\frac{\lambda}{\sigma} : \sigma \in A\right\} = P_{R|Y=y}\left\{\frac{\lambda}{\sigma} : \sigma \in A\right\}, \tag{5.3}$$

where the latter equality follows immediately from the expression of the joint distribution for (R, μ, λ, Y) derived in the proof of Theorem 3. Note that $P_{\sigma|(\mu,\lambda,y)}$ represents the conditional distribution of σ given (x, μ), since (y, λ) are the coordinates of $x - \mu$ in the particular representation chosen for $\Re^n$, i.e. $\lambda > 0$, $y \in \mathcal{Y}$ and $\lambda y = x - \mu$. Thus, from (5.3), the conditional posterior distribution of the scale parameter σ depends on the conditional distribution of R given Y, $P_{R|Y}$, and it will not be robust for $\varepsilon \in \mathcal{C}$ in (4.6). On the other hand, it is perfectly robust with respect to the choice of the marginal distribution of Y, P_Y, although this robustness is lost when we consider the marginal posterior distribution of σ given the observation (provided, of course, that it exists), since the posterior distribution of μ depends on P_Y.

However, if we are interested in certain characteristics of the posterior distribution of σ, we can still achieve robustness of such quantities in a certain subclass of $\mathcal{C}$. For instance, from (5.3) we can immediately derive that for any value $\alpha \in \Re$, the conditional posterior expectation of σ^α takes the form

$$E[\sigma^\alpha | x, \mu] = \lambda^\alpha E[R^{-\alpha} | \mu, \lambda, y] = \lambda^\alpha E[R^{-\alpha} | y],$$

where (y, λ) are the coordinates of $x - \mu$ as explained above. As the distribution of μ given x, provided that it exists, does not depend on $P_{R|Y}$ (see Theorem 3), the marginal posterior expectation of σ^α, $E[\sigma^\alpha | x]$, will only depend on $P_{R|Y}$ through its $(-\alpha)^{th}$ moment. Therefore, if we focus on the subclass of $\mathcal{C}$ such that $E[R^{-\alpha}|y]$ has a fixed value, we will obtain perfect robustness of $E[\sigma^\alpha | x]$ when ε ranges in that class.

In order for this result to be of practical interest, we would like the scale parameter σ to have some meaning in terms of sampling properties of the observables. A natural such condition would be

$$V[X | \mu, \sigma] = \sigma^2 I_n,$$

i.e. where σ^2 describes the variance of the sampling distribution. This condition translates into

$$V[\varepsilon] = V[RY] = I_n,$$

which is clearly not fulfilled in general, unless we impose some restrictions on the distribution of (Y, R). Natural conditions that lead to this situation are

$$E[Y] = 0, \quad V[Y] = bI_n, \text{ for some finite } b > 0,$$
$$E[R|Y] \text{ does not depend on } Y, \quad E[R^2|Y] = b^{-1}.$$

This implies that we can not have just any class $\mathcal{C}$ in (4.6), since we impose restrictions on the two first order moments of P_Y. Once we have chosen P_Y fulfilling these conditions, we further restrict the corresponding class $\mathcal{C}$ by only considering distributions $P_{R|Y}$ that satisfy certain moment restrictions. We shall denote the resulting class by $\mathcal{C}_0$. There are many rich classes of n-variate distributions that are compatible with these assumptions, as will be illustrated in Examples 5.3 and 5.4.

From our previous discussion follows, taking $\alpha = -2$, that the conditional expectation $E[\sigma^{-2}|x, \mu]$ is exactly the same for all choices of $\varepsilon \in \mathcal{C}_0$, and the same result holds for $E[\sigma^{-2}|x]$, the posterior expectation of the inverse variance, provided that it exists.

These results can directly be applied to the case of independent sampling from the model in (4.5), with unknown and possibly different scales, $X_i = g_i(\beta) + \sigma_i \varepsilon_i$, $i = 1, \ldots, p$, where $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d. and $\varepsilon_i = R_i Y_i$ through (2.1), under the prior structure in (5.2). If we further assume $V[\varepsilon_i] = I_n$, σ_i^2 describes the sampling variance of X_i. Following a similar reasoning as indicated above, we can derive that

$$P\left(\sigma_1 \in A_1, \ldots, \sigma_p \in A_p | x_1, \ldots, x_p, \beta\right) = \prod_{i=1}^{p} P_{R_i|Y_i=y_i}\left(\left\{\frac{\lambda_i}{\sigma_i} : \sigma_i \in A_i\right\}\right), \tag{5.4}$$

where (y_i, λ_i) are the coordinates of $x_i - g_i(\beta)$, $i = 1, \ldots, p$, in the chosen representation of $\Re^n$. Therefore, the conditional posterior means of the inverse variances, $E[\sigma_i^{-2}|x_1, \ldots, x_p, \beta]$, $i = 1, \ldots, p$, are perfectly robust when ε_i ranges in a class $\mathcal{C}_0$ as defined above. The same robustness result holds for the marginal posterior means of the inverse variances, $E[\sigma_i^{-2}|x_1, \ldots, x_p]$, $i = 1, \ldots, p$, provided that they exist, since the posterior distribution of β given $x_1, \ldots, x_p$ (whenever it is well-defined) does not depend on $P_{R_i|Y_i}$ (see Corollary 3). Of course, integrating out the common regression parameter β destroys the conditional posterior independence of the σ_i's implicit in (5.4). In particular, in the case that the marginal distribution of Y_i, P_{Y_i}, corresponds to some density function $f_2(w)$ of the polar angles $w \in \mathcal{W}$, and the prior distribution of β is given through a density $p(\beta)$ for $\beta \in \mathcal{B}$, we can obtain through (5.1)

$$E[\sigma_i^{-2}|x_1, \ldots, x_p] = E[R_i^2|y_i] \frac{\int_{\mathcal{B}} \lambda_i^{-2} p(\beta) \prod_{i=1}^{p} \|x_i - g_i(\beta)\|_2^{-n} \frac{f_2(w_i)}{s(w_i)} d\beta}{\int_{\mathcal{B}} p(\beta) \prod_{i=1}^{p} \|x_i - g_i(\beta)\|_2^{-n} \frac{f_2(w_i)}{s(w_i)} d\beta}, \tag{5.5}$$

where (y_i, λ_i), $\lambda_i > 0$, $y_i \in \mathcal{Y}$, are the coordinates of $x_i - g_i(\beta)$ as described in (2.1), $w_i = k^{-1}(y_i)$ are the angular polar coordinates as described in (2.2), and $s(w)$ comes from the Jacobian, $u^{n-1}s(w)$, of the polar transformation.

Example 5.2. l_q-spherical distributions with unitary variance

Assume the model in (4.5), $X_i = g_i(\beta) + \sigma_i \varepsilon_i$, $i = 1, \ldots, p$, where $\varepsilon_1, \ldots, \varepsilon_p$ are n-dimensional i.i.d. random vectors following an l_q-spherical distribution with unitary variance, for some fixed value of $q \in (0, \infty]$. We further consider the prior density in (5.2).

Following Osiewalski and Steel (1993), the covariance matrix of an l_q-distributed ε_i takes the form

$$V[\varepsilon_i] = c_q^{-1} E[\upsilon_q(\varepsilon_i)^2] I_n,$$

where $\upsilon_q(\cdot)$, defined in (3.2), coincides with the l_q-norm for $q \geq 1$,

$$c_\infty = \frac{3n}{n+2}, \text{ and } c_q = c_\infty \frac{\Gamma\left(1+\frac{1}{q}\right)\Gamma\left(1+\frac{n+2}{q}\right)}{\Gamma\left(1+\frac{3}{q}\right)\Gamma\left(1+\frac{n}{q}\right)} \text{ for finite } q.$$

The unitary variance assumption implies $E[\upsilon_q(\varepsilon_i)^2] = c_q$, thus defining a subset of the class of l_q-spherical distributions through fixing the second order moment of $R_i = \upsilon_q(\varepsilon_i)$.

Direct application of (5.5), with $f_2(\cdot)$ given through (3.8) with $\upsilon(\cdot) = \upsilon_q(\cdot)$, leads to the following expression for the posterior mean of the inverse variance, common to all distributions in this restricted l_q-spherical class:

$$E[\sigma_i^{-2}|x_1,\ldots,x_p] = c_q \frac{\int_{\mathcal{B}} [\upsilon_q\{x_i - g_i(\beta)\}]^{-2} p(\beta) \prod_{j=1}^p [\upsilon_q\{x_j - g_j(\beta)\}]^{-n} d\beta}{\int_{\mathcal{B}} p(\beta) \prod_{j=1}^p [\upsilon_q\{x_j - g_j(\beta)\}]^{-n} d\beta.}$$

From our previous discussion we know that this result also extends to all conditional distributions of $R_i = \upsilon_q(\varepsilon_i)$ given Y_i such that $E[R_i|Y_i]$ does not depend on Y_i and $E[R_i^2|Y_i] = c_q$. However, the case of independence between R_i and Y_i, i.e. the subset of the l_q-spherical class, seems the most interesting from a practical perspective. For instance, we know that, in the continuous case, all distributions for ε_i will then share the same isodensity sets.

Example 5.3. Linear regression with elliptical errors

Let us consider the model $X^* = D\beta + \sigma\varepsilon^*$, where ε^* is an n-variate elliptical random vector with mean zero and a known positive definite symmetric covariance matrix V, and D is a known $n \times m$ matrix of full colum rank with $n > m$. In terms of $X \equiv V^{-1/2}X^*$ and $\varepsilon \equiv V^{-1/2}\varepsilon^*$, we have a linear regression model with a spherical error vector,

$$X = V^{-1/2}D\beta + \sigma\varepsilon,$$

where $E[\varepsilon] = 0$ and $V[\varepsilon] = I_n$, which directly fits into the framework of this Subsection. Since ε follows a spherical distribution, $Y = \varepsilon/\|\varepsilon\|_2$ is uniformly distributed over the unit sphere S^{n-1} and independent of $R = \|\varepsilon\|_2$ (see Subsection 3.1.1). As the covariance matrix of the uniform distribution over S^{n-1} is $\frac{1}{n}I_n$ [see Fang *et al.* (1990, p.34)], we obtain $E[R^2] = n$ as the only restriction on the distribution of $R = \|\varepsilon\|_2$, which can also be seen from the previous example with $q = 2$.

Under the prior density $p(\beta,\sigma) = p(\beta)p(\sigma) \propto \sigma^{-1}$, the marginal posterior density for β, derived from (5.1) and (3.8), with $\upsilon(\cdot) = \|\cdot\|_2$, is the following:

$$p(\beta|x) = p(\beta|x^*) = f_S^m(\beta|n-m, \hat{\beta}, \hat{\sigma}^{-2}D'V^{-1}D),$$

which corresponds to the m-variate Student-t distribution with $n-m$ degrees of freedom, location vector $\hat{\beta} = (D'V^{-1}D)^{-1}D'V^{-1}x^*$, and precision matrix $\hat{\sigma}^{-2}D'V^{-1}D$, where $\hat{\sigma}^{-2} = (\hat{\sigma}^2)^{-1} = (n-m)\{(x^* - D\hat{\beta})'V^{-1}(x^* - D\hat{\beta})\}^{-1}$. Observe that this posterior distribution can alternatively be derived from Example 5.1 with $p = 1$.

Integrating out $E[\sigma^{-2}|x^*,\beta]$ with $p(\beta|x^*)$, leads to the posterior mean of the inverse variance

$$E[\sigma^{-2}|x^*] = n\int_{\Re^m} \frac{1}{(x^*-D\beta)'V^{-1}(x^*-D\beta)} f_S^m(\beta|n-m,\hat{\beta},\hat{\sigma}^{-2}D'V^{-1}D)d\beta = \hat{\sigma}^{-2},$$

common to all spherical distributions for ε with unitary variance.

Since $\hat{\sigma}^2$ is an unbiased estimator of σ^2, we have, for any elliptical distribution of ε^* with a fixed covariance matrix V, an interesting classical-Bayesian parallel:

$$E[\sigma^{-2}\hat{\sigma}^2|\beta,\sigma] = 1 \quad \text{and} \quad E[\hat{\sigma}^2\sigma^{-2}|x^*] = 1,$$

first noted by Osiewalski and Steel (1995).

5.2. THE SCALE MODEL

Let us consider the pure scale model introduced in (4.10), $X_i = \sigma\varepsilon_i$, $i = 1,\ldots,p$, where $\varepsilon_1,\ldots,\varepsilon_p$ are i.i.d. n-dimensional vectors (such that the distribution of ε_i given σ does not depend on this parameter), and $\sigma > 0$ is a scale parameter. We shall now present some robustness results for inference on σ, under any σ-finite prior distribution.

We first develop the theory for the case of one single vector observation. The next Theorem states the main result.

Theorem 4. *Consider the multivariate scale model $X = \sigma\varepsilon$, where ε is an n-dimensional random vector represented as $\varepsilon = RY$ through (2.1), and $\sigma > 0$ is a scale parameter. We shall factorize the joint probability distribution of (Y,R) into the marginal distribution of Y, P_Y, and the conditional distribution of R given Y, $P_{R|Y}$.*
Then, under any σ-finite prior measure for σ, the posterior distribution of σ given X, provided it exists, does not depend on P_Y.

Proof: see Appendix. •

From the proof of the Theorem follows that the key to this result is the product structure between the distributions of σ and Y, which is preserved under the one-to-one transformation from (σ,Y,R) to (σ,Y,λ), where $\lambda = \sigma R$. Thus, the conditional distribution of σ given (Y,λ), whenever it is defined, does not depend on P_Y. Note that, since (Y,λ) are the coordinates of X in the chosen representation of $\Re^n$, we can immediately derive the posterior distribution of σ given X from the conditional distribution of σ given (Y,λ).

In order for the posterior distribution of σ to exist, we require a σ-finite predictive distribution. Again from the proof of Theorem 4 we can deduce that if the following function of any Borel measurable set $C \subset \Re_+$,

$$\xi(C) = \int_0^\infty P_{R|Y}\{r : \sigma r \in C\}\mathcal{D}_\sigma(d\sigma),$$

where $\mathcal{D}_\sigma$ represents the prior distribution of σ, is a σ-finite measure on $\Re_+$, then the marginal distribution of (Y, λ), and thus the predictive distribution of X, are σ-finite. In this case, the posterior distribution of σ is defined.

In most practical situations the prior chosen for σ will either be a probability distribution, in which case the posterior distribution of σ will obviously be defined, or the standard non-informative prior, which corresponds to the density $p(\sigma) = c\sigma^{-1}$ $(c > 0)$. In this latter case, following a similar reasoning as in the beginning of Subsection 5.1.2, we can derive the following proper posterior distribution for σ

$$P_{\sigma|X=x} = P_{R|Y=y}\left\{\frac{\lambda}{\sigma} : \sigma \in A\right\},$$

where (y, λ) are the coordinates of x.

For inferential purposes on σ, we shall often be interested in a sample of several independent observations rather than in one single vector observation. The theory developed for the case of just one observation, can easily be extended to the context of repeated sampling. The following Corollary addresses this situation.

Corollary 4. *Let us consider the scale model in* (4.10), $X_i = \sigma\varepsilon_i$, $i = 1, \ldots, p$, *where* $\varepsilon_1, \ldots, \varepsilon_p$ *are i.i.d. n-dimensional vectors, represented as* $\varepsilon_i = R_i Y_i$ *through* (2.1), *and* $\sigma > 0$ *is a scale parameter. We factorize the joint probability distribution of* (Y_i, R_i) *into the marginal distribution of* Y_i, P_{Y_i}, *and the conditional distribution of* R_i *given* Y_i, $P_{R_i|Y_i}$.
Then, under any σ*-finite prior distribution for* σ, *the posterior distribution of* σ *given* $(X_1, \ldots, X_p)$, *provided it is defined, does not depend on* P_{Y_i}.

Proof: Paralleling the proof of Theorem 4, the one-to-one transformation from $(\sigma, Y_1, R_1, \ldots, Y_p, R_p)$ to $(\sigma, Y_1, \lambda_1, \ldots, Y_p, \lambda_p)$, where $\lambda_i = \sigma R_i$, preserves the product structure between the distributions of σ and $Y_1, \ldots, Y_p$. Therefore, the conditional distribution of σ given $(Y_1, \lambda_1, \ldots, Y_p, \lambda_p)$, and thus the posterior distribution of σ, provided they exist, do not depend on P_{Y_i}. •

If the prior distribution of σ is a probability measure, the posterior distribution of σ given $(X_1, \ldots, X_p)$ will always exist. However, if the prior distribution of σ is unbounded, we require the predictive distribution to be σ-finite in order to obtain a proper posterior. If the function of any Borel measurable set $C \subset \Re_+^p$ defined as

$$\rho(C) = \int_0^\infty (P_{\lambda_1|(Y_1,\sigma)} \times \ldots \times P_{\lambda_p|(Y_p,\sigma)})\{C\}\mathcal{D}_\sigma(d\sigma),$$

where $P_{\lambda_i|(Y_i,\sigma)}\{A\} = P_{R_i|Y_i}\{r_i : \sigma r_i \in A\}$, for each measurable set $A \subset \Re_+$, is a σ-finite measure on $\Re_+^p$, then the marginal distribution of $(Y_1, \lambda_1, \ldots, Y_p, \lambda_p)$, or, equivalently, the predictive distribution of $(X_1, \ldots, X_p)$, is σ-finite. In this case, the posterior distribution of σ given $(X_1, \ldots, X_p)$ is well-defined.

Finally, let us examine the situation in which both $P_{R_i|Y_i}$ and the prior distribution of σ are given through density functions $f_1(r_i; y_i)$ and $p(\sigma)$, respectively. The posterior

distribution of σ given $(X_1, \ldots, X_p)$ will now be characterized through the density

$$p(\sigma|x_1, \ldots, x_p) \propto p(\sigma) \prod_{i=1}^{p} \frac{1}{\sigma} f_1 \left(\frac{\lambda_i}{\sigma}; y_i \right),$$

where (y_i, λ_i) are the coordinates of x_i, $i = 1, \ldots, p$, which requires

$$\int_0^\infty p(\sigma) \prod_{i=1}^{p} \frac{1}{\sigma} f_1 \left(\frac{\lambda_i}{\sigma}; y_i \right) d\sigma < \infty$$

in order to be proper. Clearly, this density does not depend on the distribution of Y_i, P_{Y_i}, as we already knew from Corollary 4.

From the discussion in this Subsection we conclude that in the pure scale model in (4.10), inference on σ, whenever it can be conducted, does not depend on P_{Y_i}. Therefore, inference on σ is perfectly robust in the class of distributions for the error term $\varepsilon_i = R_i Y_i$ that share the same conditional distribution $P_{R_i|Y_i}$.

Note that, when we discussed robustness of classical inferences for the same scale model in (4.10), we considered pivotal quantities that are only functions of σ and $(\lambda_1, \ldots, \lambda_p)$, the radial coordinates of each of the observations, and, therefore, their sampling distribution only depends on the marginal distribution of R_i, P_{R_i}, derived from $P_{(Y_i,R_i)}$. This immediately leads to robustness of classical inference on σ with respect to the choice of the conditional distribution of Y_i given R_i. On the other hand, Corollary 4 states robustness of Bayesian inference on σ with respect to the choice of the marginal distribution of Y_i, P_{Y_i}. Thus, our classical and Bayesian robustness results refer to opposite factorizations of the joint distribution of (Y_i, R_i). In addition, Corollary 4 tells us that drawing inferences only on the basis of $(\lambda_1, \ldots, \lambda_p)$, the radial coordinates of the observations, while discarding the data on $(Y_1, \ldots, Y_p)$, leads to a loss of relevant information about σ if R_i and Y_i are not independent. Of course, if we are interested in distributions of the error term $\varepsilon_i = R_i Y_i$ that impose independence between R_i and Y_i, $P_{R_i|Y_i}$ no longer depends on Y_i, both factorizations of $P_{(Y_i,R_i)}$ coincide and we obtain a parallelism between classical and Bayesian results. In this case, both classical and Bayesian inference on scale are completely robust when ε_i ranges in a class $\mathcal{R}$ in (3.12). Robustness of classical inference in the l_q-anisotropic class, defined in Subsection 3.2, was illustrated in Example 4.3. We now present a parallel Bayesian example.

Example 5.4. Independent sampling from an l_q-anisotropic scale model

As in Example 4.3, the p vector observations are generated from the scale model (4.10) $X_i = \sigma \varepsilon_i$, $i = 1, \ldots, p$, where each element of each vector ε_i is independently sampled from the same exponential power distribution with some $q \in (0, \infty]$. For $q = \infty$, this corresponds to a uniform sampling distribution on the n-rectangle $(-\sigma, \sigma)^n$ for each X_i.

For the scale parameter σ, we assume a proper prior, corresponding to

$$b\sigma^{-q} \sim \chi^2_{2a/q}, \text{ for some } a, b > 0, \text{ for } q \in (0, \infty),$$

and

$$b_* \sigma^{-1} \sim \text{Beta}(a_*, 1), \text{ for some } a_*, b_* > 0, \text{ for } q = \infty.$$

Since we have a proper prior, the posterior distribution of σ will clearly exist. In order to calculate the latter, we recall that for $R_i = v_q(\varepsilon_i)$ the distribution given $Y_i = \varepsilon_i / v_q(\varepsilon_i)$ is characterized by

$$f_1(r_i; y_i) = n \left\{ \Gamma\left(1 + \frac{n}{q}\right) \right\}^{-1} 2^{-n/q} r_i^{n-1} \exp(-r_i^q/2), \text{ for finite } q,$$

and by its limit

$$f_1(r_i; y_i) = n r_i^{n-1} I_{(0,1)}(r_i), \text{ for } q = \infty.$$

Direct calculations lead to the posterior distribution of σ, described by

$$\left(b + \sum_{i=1}^{p} \{v_q(x_i)\}^q \right) \sigma^{-q} | (x_1, \ldots, x_p) \sim \chi^2_{2(a+pn)/q}, \text{ when } 0 < q < \infty,$$

and by

$$\max\{b_*, \max_{j=1,\ldots,n; i=1,\ldots,p} |x_{ji}|\} \sigma^{-1} | (x_1, \ldots, x_p) \sim \text{Beta}(a_* + np, 1), \text{ for infinite } q,$$

where x_{ji}, $j = 1, \ldots, n$, represent the n components of the i^{th} observation. Applying the theory described above, we obtain that whenever the distribution of R_i given Y_i remains the same distribution as assumed here, i.e. when ε_i ranges in the entire l_q-anisotropic class for given q, the posterior distribution of σ is unaffected.

In the limit as a, b, a_* and b_* all tend to zero, the kernels of our prior densities of σ tend to $p(\sigma) \propto \sigma^{-1}$, and the posterior distribution of σ tends to the sampling distribution for the pivots in Example 4.3, considered as a function of σ.

6. 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, once again, 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 with scale and pure scale models to analyze the robustness of classical inference in certain classes of distributions. We do not actually investigate particular inference procedures, but rather analyze this robustness through distribution-freeness of pivots in these classes.

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 $t(Z_1, \ldots, Z_p)$

are distribution-free if and only if their distribution is not affected by rescaling each Z_i. Applying this result in the context of the regression model under i.i.d. sampling up to a scale, $X_i = g_i(\beta) + \sigma_i \varepsilon_i$, $i = 1, \ldots, p$, we find that a function $t(X_1 - g_1(\beta), \ldots, X_p - g_p(\beta))$ is distribution-free for $\varepsilon_i \in \mathcal{S}$ if and only if it is a pivotal quantity, i.e. its distribution does not depend on $(\beta, \sigma_1, \ldots, \sigma_p)$. This immediately leads to robustness of classical inference on the regression parameter based on such pivots.

For the, practically less interesting, classes where we fix the distribution of R, obtaining a characterization of invariance in terms of a set of transformations on the observables is less immediate. Clearly, functions that only depend on R will be distribution-free in such classes. In fact, inference procedures based on them will be unaffected by the conditional distribution of Y given R. For the pure scale model this situation arises naturally.

Another major goal of the paper was to investigate the robustness of Bayesian inference for the same type of models. In addition, we set out to compare these results with the classical findings. Our Bayesian robustness results are not derived from distribution theory results applied to the sampling distribution, but rather follow from the use of measure theory on the joint space of observables and parameters. This gives the results a somewhat different nature. However, interesting parallels can be uncovered.

For the regression model mentioned above, under 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_i 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 the error term ε_i that impose independence between R and Y, thus leading to robustness of both classical and Bayesian inference in the same classes of distributions. 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 inference under both paradigms that each observation has its own scale parameter. However, the reasons are different: the sampling theory results no longer lead to a characterization of the distribution-free quantities under a common scale and, in particular, not all pivotal quantities are distribution free for $\varepsilon_i \in \mathcal{S}$, whereas Bayesian robustness disappears under a common scale. Of course, the Bayesian analysis always has to formally incorporate the extra information that all scales are equal (through specifying a prior on the common scale parameter), whereas this is often not the case in a classical framework.

If we wish to conduct inference on the mean of the precision of the observables in a regression model, we find that a Bayesian analysis will lead to the same posterior mean in a subclass of $\mathcal{C}$. Basically, this posterior mean will only depend on the distribution of R given Y through the second order moment $E[R^2|Y]$, so fixing the latter will naturally lead to robustness. In the special case of a linear regression model with elliptical errors, this finding also has a classical counterpart (which, however, does not rely on ellipticity).

We stress that both classical and Bayesian robustness results in the regression model ultimately depend on the presence of a scalar scale parameter. It is this parameter that naturally leads to distribution-free pivotal quantities in the classical framework, and in

a Bayesian setting integrating out this scale parameter (with a particular σ-finite prior) generates our robustness results. Introducing a location vector into the model, or other additions like e.g. a parametric covariance structure, is only important for inference purposes on the associated parameters, but does not drive any of the invariance results. Once we have a scale parameter, all our robustness results hold given these additional parameters, which are often the parameters of interest.

Bayesian robust inference on the scale parameter in the pure scale model is seen to hold for any prior. However, in contrast to the classical result, we now have robustness with respect to the marginal distribution of Y. The Bayesian perspective clearly shows that inference on the scale based on only R involves a loss of information, unless R and Y are independent. Again, under independence between R and Y, we have a perfect parallel between sampling theory inference based on any function of R and Bayesian inference.

An important issue in the Bayesian paradigm 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 the standard location-scale model, $X_i = \mu + \sigma_i \varepsilon_i$, $i = 1, \ldots, p$, under the prior on the scales $p(\sigma_1, \ldots, \sigma_p) \propto \prod_{i=1}^{p} \sigma_i^{-1}$ that assures robustness when $\varepsilon_i \in \mathcal{C}$, typically does not allow for a proper posterior of the location parameter μ. In order to conduct inference on the location, we usually require a reparameterization in terms of a lower dimensional regression parameter β.

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 models treated here. In addition, we think our understanding of statistical inference procedures in such models is furthered by contrasting these classical results with their Bayesian counterparts.

APPENDIX

Proof of Theorem 1

First, assume (4.2), which is equivalent to $t(Z) \stackrel{d}{=} t(Z^*)$, for all $Z, Z^* \in \mathcal{MS}$. From the definition of $\mathcal{MS}$ in (4.1) 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 choice of $\alpha_1, \ldots, \alpha_p > 0$. Thus, we immediately obtain (4.3).

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 can apply (4.3) with $\alpha_i = r_i$ to obtain that

$$F_{t(r_1Y_1,\ldots,r_pY_p)}(x) = F_{t(Y_1,\ldots,Y_p)}(x).$$

Substituting the latter inside the integral, the result (4.2) follows. •

Proof of Theorem 2

Here we shall only prove that (ii) implies (i), as all other results are easily derived. For $Z = (Z_1, \ldots, Z_p) \in \mathcal{MR}$, we consider the distribution function of $t(Z)$:

$$F_{t(Z_1,\ldots,Z_p)}(x) = F_{t(R_1 k_1(W_1),\ldots,R_p k_p(W_p))}(x)$$
$$= \int_{\mathcal{W}^p} F_{t(R_1 k_1(W_1),\ldots,R_p k_p(W_p))|(W_1,\ldots,W_p)=(w_1,\ldots,w_p)}(x) F_{(W_1,\ldots,W_p)}(dw_1, \ldots, dw_p).$$

Due to the independence between $(R_1, \ldots, R_p)$ and $(W_1, \ldots, W_p)$ imposed in $\mathcal{MR}$, we obtain that

$$F_{t(R_1 k_1(W_1),\ldots,R_p k_p(W_p))|(W_1,\ldots,W_p)=(w_1,\ldots,w_p)}(x) = F_{t(R_1 k_1(w_1),\ldots,R_p k_p(w_p))}(x),$$

to which (ii) can be applied directly. Thus, the result follows. •

Proof of Theorem 3

We shall derive the joint distribution of $(X - \mu, \mu)$, from which the distribution of (X, μ) immediately follows.
First note that

$$(X - \mu, \mu) = (\sigma\varepsilon, \mu) = (\sigma R Y, \mu) = (\lambda Y, \mu), \text{ where } \lambda = \sigma R.$$

By assumption, the distribution of (μ, σ, Y, R), denoted by $\mathcal{D}_{(\mu,\sigma,Y,R)}$, is the product measure

$$\mathcal{D}_{(\mu,\sigma,Y,R)} = \mathcal{D}_\mu \times \mathcal{D}_\sigma \times P_{(Y,R)},$$

where $\mathcal{D}_\mu$ is the prior measure of μ, $\mathcal{D}_\sigma$ 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 (μ, σ, Y, R) to (μ, λ, Y, R), where $\lambda = \sigma R$, the distribution of (μ, λ, Y, R) will be given by $\mathcal{D}_{(\mu,\lambda,Y,R)} = \mathcal{D}_\mu \times \mathcal{D}_{(\lambda,Y,R)}$, 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

$$f^{-1}(C)_{(y,r)} = \{\sigma : (\sigma, y, r) \in f^{-1}(C)\} = \{\sigma : (\sigma r, y, r) \in C\} = \{\sigma : \sigma r \in C_{(y,r)}\},$$

where $C_{(y,r)} = \{\sigma : (\sigma, y, r) \in C\}$.
Due to the invariance property of $\mathcal{D}_\sigma$ under the group of transformations $\{r\sigma : r > 0\}$, it is immediately derived 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

$$\mathcal{D}_{(\lambda,Y,R)}(C) = \mathcal{D}_{(\sigma,Y,R)}\{f^{-1}(C)\} = \mathcal{D}_{(\sigma,Y,R)}(C),$$

which implies

$$\mathcal{D}_{(\mu,\lambda,Y,R)} = \mathcal{D}_\mu \times \mathcal{D}_\lambda \times P_{(Y,R)}, \qquad \text{where } \mathcal{D}_\lambda = \mathcal{D}_\sigma.$$

In order to recuperate the distribution of $(X - \mu, \mu) = (\lambda Y, \mu)$, we would be interested in the distribution of (μ, λ, Y). By definition of the marginal distribution, the measure of any set A will be given by

$$\mathcal{D}_{(\mu,\lambda,Y)}(A) = \mathcal{D}_{(\mu,\lambda,Y,R)}\{A \times (0,\infty)\} = (\mathcal{D}_\mu \times \mathcal{D}_\sigma \times P_Y)\{A\}P_{R|Y}\{(0,\infty)\}.$$

As $P_{R|Y}\{(0,\infty)\} = 1$ for all possible choices of probability measures $P_{R|Y}$, the distribution of (μ, λ, Y) does not depend on $P_{R|Y}$. Therefore, the distributions of $(\lambda Y, \mu) = (X - \mu, \mu)$ and (X, μ) do not depend on $P_{R|Y}$ either. •

Proof of Proposition 2

We can always find an open set, say $A_0 \subset \Re^n$, with positive Lebesgue measure, such that $p(\beta) > K^*$ for all $\beta \in A_0$ and some constant $K^* > 0$. Now let us consider any sample $(x_1, \ldots, x_p)$ such that $x_1 \in A_0$, and choose a constant $\delta > 0$ such that the n-ball $B(x_1, \delta) = \{\beta \in \Re^n : \|x_1 - \beta\|_2 < \delta\}$ is contained in A_0.

Using the triangle inequality, we obtain for $i = 2, \ldots, p$, $\|x_i - \beta\|_2 \leq \|x_i - x_1\|_2 + \|x_1 - \beta\|_2$. From (5.1) it can be seen that

$$p(x_1, \ldots, x_p) \geq cK^* \int_{B(x_1,\delta)} \|x_1 - \beta\|_2^{-n} \prod_{i=1}^{p} \frac{f_2\left\{h^{-1}\left(\frac{x_i-\beta}{\|x_i-\beta\|_2}\right)\right\}}{s\left\{h^{-1}\left(\frac{x_i-\beta}{\|x_i-\beta\|_2}\right)\right\}} d\beta \prod_{i=2}^{p} (\|x_i - x_1\|_2 + \delta)^{-n}.$$

Our assumption that $f_2(w)/s(w) > K > 0$ for all $w \in \mathcal{W}$, leads to

$$p(x_1, \ldots, x_p) \geq cK^* \int_{B(x_1,\delta)} K^p \|x_1 - \beta\|_2^{-n} d\beta \prod_{i=2}^{p} (\|x_i - x_1\|_2 + \delta)^{-n}.$$

Making the change of variables from β to the polar coordinates of $x_1 - \beta$, we can immediately verify that the latter integral is not finite.

Since this reasoning holds for any sample $(x_1, \ldots, x_p) \in A_0 \times \Re^{n\times(p-1)}$, which obviously has non-zero Lebesgue measure, the predictive distribution of $(X_1, \ldots, X_p)$ is not σ-finite.

Proof of Theorem 4

First note that

$$(\sigma, X) = (\sigma, \sigma R Y) = (\sigma, \lambda Y), \quad \text{where } \lambda = \sigma R.$$

Observe that (Y, λ) are the coordinates of X in the chosen representation of $\Re^n$ and, thus, from the conditional distribution of σ given (Y, λ), we can immediately derive the posterior distribution of σ given X.
By hypothesis, the joint distribution of (σ, Y, R), denoted by $\mathcal{D}_{(\sigma,Y,R)}$, can be factorized as

$$\mathcal{D}_{(\sigma,Y,R)} = \mathcal{D}_\sigma \times P_{(Y,R)} = \mathcal{D}_\sigma \times P_Y \times P_{R|Y},$$

where $\mathcal{D}_\sigma$ is the σ-finite prior distribution of σ, and $P_{(Y,R)}$ represents the joint probability distribution of (Y, R), which we further factorize into the marginal distribution of Y, P_Y, and the conditional distribution of R given Y, $P_{R|Y}$.

If we now consider the one-to-one transformation from (σ, Y, R) to (σ, Y, λ), where $\lambda = \sigma R$, we obtain that for any measurable set $A \times B$, the measure of the event $\{\sigma \in A \text{ and } (Y, \lambda) \in B\}$ is given by

$$\begin{aligned} \mathcal{D}_{(\sigma,Y,\lambda)}\{A \times B\} &= \mathcal{D}_{(\sigma,Y,R)}\{(\sigma, y, r) : (\sigma, y, \sigma r) \in A \times B\} \\ &= \int_{\mathcal{Y}} \int_A P_{R|Y=y}\{r : \sigma r \in B_y\} \mathcal{D}_\sigma(d\sigma) P_Y(dy), \end{aligned} \tag{A.1}$$

where $B_y = \{s > 0 : (y, s) \in B\}$. From the latter expression follows that

$$\mathcal{D}_{(\sigma,Y,\lambda)} = \mathcal{D}_\sigma \times P_Y \times P_{\lambda|(Y,\sigma)},$$

where, for each measurable set C,

$$P_{\lambda|(Y,\sigma)}\{C\} = P_{R|Y}\{r : \sigma r \in C\}.$$

In order for the conditional probability distribution of σ given (Y, λ) to be defined, we require that the marginal distribution of (Y, λ) is σ-finite. This marginal distribution, derived from $(A.1)$, takes the form

$$\mathcal{D}_{(Y,\lambda)}(B) = \mathcal{D}_{(\sigma,Y,\lambda)}\{(0, \infty) \times B\} = \int_{\mathcal{Y}} \int_0^\infty P_{R|Y=y}\{r : \sigma r \in B_y\} \mathcal{D}_\sigma(d\sigma) P_Y(dy), \tag{A.2}$$

where B_y was defined above, for each measurable set B. From $(A.1)$ and $(A.2)$ immediately follows that the conditional distribution of σ given (Y, λ), whenever it exists, does not depend on P_Y.

REFERENCES

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.

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

Dickey, J.M. and Chen, C.H. (1985), "Direct subjective-probability modelling using ellipsoidal distributions", in *Bayesian Statistics 2,* eds. Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.F.M., Amsterdam: North-Holland, pp. 157-182.

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) "Modelling and inference with υ-spherical distributions", *Journal of the American Statistical Association*, forthcoming.

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.

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

Osiewalski, J. and Steel, M.F.J. (1995), "Posterior moments of scale parameters in elliptical sampling models", in *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner,* eds. Berry, D., Chaloner, K. and Geweke J., New York: Wiley, forthcoming.

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

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics,* New York: Wiley.