Reference Priors
for Non-Normal Two-Sample Problems

By Carmen Fernández and Mark F.J. Steel[1]

CentER for Economic Research and Department of Econometrics
Tilburg University, 5000 LE Tilburg, The Netherlands

ABSTRACT

The reference prior algorithm (Berger and Bernardo, 1992) is applied to location-scale models with any regular sampling density. A number of two-sample problems is analyzed in this general context, extending the difference, ratio and product of Normal means problems outside Normality, while explicitly considering possibly different sizes for each sample. Since the reference prior turns out to be improper in all cases, we examine existence of the resulting posterior distribution and its moments under sampling from scale mixtures of Normals. In the context of an empirical example, it is shown that a reference posterior analysis is numerically feasible and can display some sensitivity to the actual sampling distributions. This illustrates the practical importance of questioning the Normality assumption.

Keywords: Behrens-Fisher problem; Fieller-Creasy problem; Gibbs sampling; Jeffreys' prior; Location-scale model; Posterior existence; Product of means; Scale mixtures of Normals; Skewness.

1. INTRODUCTION

The search for a standard "non-informative" prior distribution, to formally express prior ignorance or for the purpose of scientific reporting started in earnest in Jeffreys (1961). The principle introduced in Jeffreys (1961) has gained widespread acceptance for models with only one parameter. However, in the presence of multiple parameters, various alternative approaches have been suggested. Jeffreys himself (1961, p.182) considers a modification of his rule for the cases where "a previous judgement of irrelevance" seems reasonable. The resulting prior will be denoted as the "independence Jeffreys' prior" in the sequel.

A formal methodology for multiparameter models was introduced in Bernardo (1979) on the basis of information theory arguments and distinguishing between parameters of interest and nuisance parameters. This "reference prior" algorithm was further developed and defined in Berger and Bernardo (1992). They introduce and discuss some technical refinements, such as a nested sequence of compact sets (hereafter denoted by $\{\Theta^l\}$), converging to the entire parameter space when the latter is non-compact. The reference prior is then derived by first considering Θ^l and afterwards taking a limit on l. In addition, in

[1] Corresponding author.

Useful comments from José M. Bernardo, Jacek Osiewalski and an anonymous Referee are gratefully acknowledged. The first author holds a Research Training Grant (ERBFMBICT 961021) under the Training and Mobility of Researchers Programme, financed by the European Commission.

the case of more than two parameters, they consider a finer grouping of parameters than the one that merely separates the parameter of interest from the nuisance parameters. The choice of the sequence of sets, as well as the choice and order of the parameter groups are found to potentially influence the form of the reference prior. On the basis of their experience with various models, Berger and Bernardo (1992) state that usually the choice of the sequence $\{\Theta^l\}$ does not matter, and they recommend using a separate group for each parameter. For the ordering of the groups, however, they give no strict guidelines, other than ordering according to "inferential importance". Thus, the form of the reference prior can be subject to a number of essentially arbitrary decisions.

In this paper, we shall consider the Berger and Bernardo reference prior. Some recent publications in this area are, among others, Sun and Ye (1995), Datta and Ghosh (1995) and Clarke (1996). Alternative noninformative prior distributions are proposed in *e.g.* Tibshirani (1989), Ghosh and Mukerjee (1992) and Clarke and Wasserman (1993). Kass and Wasserman (1996) give a comprehensive overview of such prior selection rules.

The present paper focuses on the case of two independent samples from multivariate location-scale models and inference concerning some function of both locations. In particular, we aim to extend the distributional assumptions under which reference priors have been derived from the restrictive Normal sampling to any combination of regular sampling densities. Section 2 presents some general results. An interesting feature is that the ratio of the sample sizes may enter the reference prior. Due to the presence of more than two parameters, grouping and ordering the parameters becomes an issue. Throughout the paper, we shall follow the recommendation of Berger and Bernardo (1992) to put each parameter in a separate group. If, in addition, the scales are ranked last, the reference prior is seen to be the product of the independence Jeffreys' prior and some function of the locations. The latter function can depend on the sampling distributions, the ratio of the sample sizes, and the choices of the parameter of interest and the sequence $\{\Theta^l\}$. Thus, whereas the independence Jeffreys' prior is the same for any two-sample problem, the form of the reference prior may vary according to some features of the problem. In particular, its dependence on the parameter of interest is an intrinsic characteristic of this method. Different choices of the parameter of interest define different problems, and in Sections 3-5 we discuss in detail the difference, ratio and product of locations. In this way, we respectively generalize the well-known Behrens-Fisher, Fieller-Creasy and product of means problems, previously examined in the reference prior literature under the assumption of two Normal samples (see Liseo, 1992, 1993; Bernardo, 1977; Berger and Bernardo, 1989; Bernardo and Smith, 1994), to any two regular continuous location-scale models with possibly different sample sizes. In addition, we also treat the product of locations with unknown equal scale.

Since the reference priors thus derived are all improper, existence of the posterior distribution becomes an issue. Thus, we complement our results with necessary and sufficient conditions for propriety of the posterior under sampling from the practically useful class of scale mixtures of Normals. Following Florens, Mouchart and Rolin (1990), we have a well-defined conditional distribution of the parameters given the observables (*i.e* a proper posterior distribution) whenever the marginal distribution of the observables is σ-finite. The latter, however, does not exclude the possibility that its density becomes infinite in a set of Lebesgue measure zero, which would preclude posterior inference for samples in that set. The origins of this potential danger reside in the general "incompatibility" between continuous sampling models and point observations, and is not specific to the use of improper priors. This issue, however, is outside the scope of the present paper.

Section 6 contains an empirical illustration of the difference and ratio of locations, using a clinical data set presented in Karpatkin, Porges and Karpatkin (1981). We contrast different sampling assumptions and illustrate the feasibility of a reference analysis under a variety of sampling distributions.

Some main conclusions are summarized in Section 7.

Throughout the paper, we assume the sufficient regularity conditions for asymptotic Normality of the likelihood function in DeGroot (1970, Ch.10), which implies that the reference prior can be derived solely on the basis of the information matrix (Section 2.3 of Berger and Bernardo, 1992, and Proposition 6.30 of Bernardo and Smith, 1994). For the particular sampling distributions considered in some detail in this paper, *i.e.* Student-t and Skewed Exponential Power (with $q > 1$) (see Appendix B), we have checked that asymptotic Normality of the likelihood function holds. For reference priors in non-regular models, see Ghosal (1997).

Finally, we shall use $p(\cdot)$ to denote probability density functions on observables, whereas the notation $\pi(\cdot)$ shall be reserved for parameters. Proofs are sketched in Appendix A.

2. TWO SAMPLES FROM GENERAL LOCATION-SCALE MODELS

We consider m independent and identically distributed (i.i.d.) replications from a general r-variate location-scale model, with density function

$$p(x|\alpha,\sigma) = \sigma^{-r} f\{\sigma^{-1}(x-\alpha)\}, \tag{2.1}$$

where $x \in \Re^r$, $\alpha \in \Re^r$ is a location parameter, $\sigma > 0$ is a scale parameter and $f(\cdot)$ is a probability density function (p.d.f.) in $\Re^r$. A second independent sample of n i.i.d. observations is assumed to be generated from the model

$$p(y|\beta,\xi) = \xi^{-s} g\{\xi^{-1}(y-\beta)\}, \tag{2.2}$$

where $y \in \Re^s$, $\beta \in \Re^s$ is a location parameter, $\xi > 0$ is the scale and $g(\cdot)$ is a p.d.f. in $\Re^s$. Implicitly, we shall assume throughout that the regularity conditions for asymptotic Normality mentioned in the Introduction hold for both (2.1) and (2.2).

Our parameter of interest, denoted by θ, will be some function of the locations α and β, and can be either uni- or multi-dimensional depending on the context. Under Normal sampling, three well-known cases that have been previously addressed in the reference prior literature are the difference of means (Behrens-Fisher problem), the ratio of means with $\sigma = \xi$ (Fieller-Creasy problem) and the product of means with known σ and ξ. In this paper, we mainly concentrate on analyzing the extensions of these classical two-sample problems to the context of two general location-scale models as in (2.1) and (2.2), without the Normality assumption.

Observe that we explicitly allow for any sample sizes, m and n, possibly different. It is well-known that in the one-sample case, sample size does not affect the reference prior. The basic experiment, upon which the reference prior is based, is then just one replication. In the two-sample situation, however, relative sample size m/n enters the information matrix in a non-trivial way. The basic experiment in this case requires fixing relative sample size, and thus the reference prior can depend on the ratio m/n. However, it will not depend on the number of i.i.d. replications from the basic experiment, parallelling the result for one sample in that sense.

First, we present a general result on reference priors for two-sample problems. Throughout the paper, we will reparameterize (α, β) into (θ, ρ), where θ is the parameter of interest and ρ some nuisance parameter. We shall always assume a Cartesian product structure between (θ, ρ), σ and ξ in Θ^l; this will be denoted as

$$\Theta^l = \Theta^l_{\theta,\rho} \times \Theta^l_{\sigma} \times \Theta^l_{\xi}.$$

We now distinguish between the case of unknown but equal scales (*i.e.* $\sigma = \xi$) and the case where σ and ξ are unknown and potentially different.

Proposition 1. *Consider two independent samples of m i.i.d. replications from* (2.1) *and n i.i.d. replications from* (2.2).
(i) <u>*Equal scales:*</u> *We assume $\sigma = \xi$ and take three groups ordered as $\{\theta, \rho, \sigma\}$. Then the reference prior in the original parameterization is*

$$\pi(\alpha, \beta, \sigma) \propto \sigma^{-1} R(\alpha, \beta), \tag{2.3}$$

for some non-negative function $R(\cdot)$, the form of which may depend on the choice of (θ, ρ), as well as on $f(\cdot)$, $g(\cdot)$, m/n and $\{\Theta^l_{\theta,\rho}\}$.
(ii) <u>*Potentially different scales:*</u> *With four groups ordered as $\{\theta, \rho, \sigma, \xi\}$ or $\{\theta, \rho, \xi, \sigma\}$, the reference prior is*

$$\pi(\alpha, \beta, \sigma, \xi) \propto \sigma^{-1} \xi^{-1} Q(\alpha, \beta), \tag{2.4}$$

for some non-negative function $Q(\cdot)$, the form of which may depend on (θ, ρ), as well as on $f(\cdot)$, $g(\cdot)$, m/n and $\{\Theta^l_{\theta,\rho}\}$.

Remark that the independence Jeffreys' prior would always be $\pi(\alpha, \beta, \sigma) \propto \sigma^{-1}$ for the case $\xi = \sigma$ [see $(A.1)$ in Appendix A], and $\pi(\alpha, \beta, \sigma, \xi) \propto \sigma^{-1}\xi^{-1}$ for possibly different scales [from $(B.1)$ in Appendix B]. Our results in Proposition 1 add an additional factor, $R(\cdot)$ or $Q(\cdot)$, which depends on a number of features of the problem. In line with the underlying motivation for the reference prior, the choice of the parameter of interest plays a crucial role in the form of the reference prior (as we shall see in the next three sections), whereas it leaves the Jeffreys' prior totally unaffected.

3. THE DIFFERENCE OF LOCATIONS

In the Behrens-Fisher problem (see *e.g.* Fisher, 1956, chap. 4) the issue is to conduct inference on the difference of the means of two independent Normal samples with unknown and possibly different variances. The reference prior for this problem, which corresponds to choosing for $f(\cdot)$ in (2.1) and for $g(\cdot)$ in (2.2) standard Normal p.d.f.'s and $r = s$, was derived in Liseo (1992) for $r = s = 1$. In this section, we examine the problem of the difference of locations when we drop the assumption of Normality and we consider instead any two location-scale models with any value for $r = s$ and any two sample sizes m and n. We thus take the parameter of interest to be $\theta = \alpha - \beta$, and we shall choose as nuisance parameters β, σ and ξ. Our main result is summarized in the following theorem.

Theorem 1. *Consider four groups in the order* $\{\theta, \beta, \sigma, \xi\}$ *or* $\{\theta, \beta, \xi, \sigma\}$, *with* $\Theta^l_{\theta,\beta} = \Theta^l_\theta \times \Theta^l_\beta$ *or* $\Theta^l_{\theta,\beta}$ *corresponding to rectangles for* (α, β) *such that* $\alpha_j \in [-a_j(l), a_j(l)]$, $\beta_j \in [-b_j(l), b_j(l)]$, *where* $\{a_j(l)\}$ *and* $\{b_j(l)\}$ *are increasing sequences of positive numbers with* $\lim_{l\to\infty}\{a_j(l)/b_j(l)\} = 1$ *for all* $j = 1, \ldots, r$ *(*α_j *and* β_j *denote the* j^{th} *component of* α *and* β*, respectively). Then we obtain as the reference prior*

$$\pi(\theta, \beta, \sigma, \xi) = \pi(\alpha, \beta, \sigma, \xi) \propto \sigma^{-1}\xi^{-1}, \tag{3.1}$$

for any choice of $f(\cdot)$ *in* (2.1) *and* $g(\cdot)$ *in* (2.2), *with any* $r = s$, m *and* n.

Remark that the Normality assumption plays no role whatsoever when $\Theta^l_{\theta,\beta} = \Theta^l_\theta \times \Theta^l_\beta$. In this problem, however, one could find it more natural to focus on the original parameterization. From Theorem 1 we see that if $\Theta^l_{\theta,\beta}$ corresponds to a natural sequence of rectangles for (α, β), (3.1) is still the reference prior. Clearly, relative sample size does not affect the reference prior in (3.1) either. In terms of Proposition 1 (ii), (3.1) corresponds to $Q(\alpha, \beta) = 1$, regardless of $f(\cdot)$, $g(\cdot)$ and m/n, and thus coincides with the independence Jeffreys' prior.

Since we now have more than one nuisance parameter, their ordering becomes an issue. Theorem 1 addresses the situation where the scales are ranked last [as was the case in Proposition 1 (ii)]. If at least one of the scales precedes β, then with the same choices of $\{\Theta^l\}$ we can still obtain (3.1) under the following condition:

Proposition 2. *If* $f(\cdot)$ *and* $g(\cdot)$ *both lead to a block diagonal information matrix between location and scale [i.e.* $b(f) = b(g) = 0$ *in* $(B.1)$*], then* (3.1) *is the reference prior (with four groups) under any ordering of the nuisance parameters.*

The condition of Proposition 2 is assured whenever we have axial symmetry (*e.g.* sphericity, which includes the class of scale mixtures of Normals described in (3.2), or l_q-sphericity as defined in Osiewalski and Steel, 1993), but also holds in other cases such as Skewed Exponential Power distributions, as explained in Appendix B.

When block diagonality fails to hold for at least one of the information matrices, we typically lose the form of the reference prior in (3.1) if one or both scales precede β. As mentioned in Berger and Bernardo (1992), the ordering of the nuisance parameters is essentially arbitrary and different orderings can lead to different forms of the reference prior.

As a result of the particular product structure of the reference prior in (3.1) and the independence between both samples, (α, σ) and (β, ξ) are independent a posteriori, and propriety of the posterior distribution holds if and only if it holds for each of the samples separately. In addition, positive-order posterior moments of θ exist if and only if the corresponding posterior moments of α and β exist. As an important example, we consider sampling from scale mixtures of Normals, which corresponds to the following choice of

$f(\cdot)$:

$$f(z) = \int_0^\infty \left(\frac{\lambda}{2\pi}\right)^{r/2} \exp\left(-\frac{\lambda}{2} z'z\right) dP_\lambda, \tag{3.2}$$

where $z \in \Re^r$ and P_λ is any probability distribution for the mixing variable λ in $\Re_+$. This is a rich class, which contains as leading examples the r-variate Normal and Student-t distributions.

Proposition 3. *Combining m i.i.d. replications from* (2.1) *with $f(\cdot)$ as in* (3.2) *with the prior $\pi(\alpha, \sigma) \propto \sigma^{-1}$, leads to a well-defined conditional distribution of the parameters given the observables if and only if $m \geq 2$. Positive-order posterior moments of the components of α exist up to the order $r(m-1)$ (not including).*

In a multivariate framework $(r, s > 1)$, we could also reparameterize α and β in terms of lower-dimensional vectors γ and δ, respectively, and focus our interest on $\gamma - \delta$. Note that now we can allow for x and y to have different dimensions (*i.e.* $r = s$ is no longer imposed) as long as both γ and δ share some lower dimension p. In the extreme case of common location, $\alpha = \gamma\iota_r$ and $\beta = \delta\iota_s$ with $p = 1$ and ι_q denoting a $q \times 1$ vector of ones. Then, using the four groups $\gamma - \delta$, δ, σ, ξ, the results in Theorem 1 and Proposition 2 directly apply, defining $\theta = \gamma - \delta$ and replacing α and β by γ and δ respectively.

4. THE RATIO OF LOCATIONS

As before, we shall consider m i.i.d. observations from (2.1) and n i.i.d. observations from (2.2) under independence between both samples. We assume $r = s = 1$, *i.e.* univariate observations. The focus of interest is now the ratio of locations: $\theta = \alpha/\beta$. Traditionally, this problem has been posed in the context of Normality and equal unknown variances σ^2. The seminal discussion of this issue can be found in Fieller (1954) and Creasy (1954), who both propose different fiducial solutions to this problem.

Bernardo (1977) derives the reference prior in this Normal case, and obtains for θ being the parameter of interest:

$$\pi(\alpha, \beta, \sigma) \propto \sigma^{-1}(\alpha^2 + \beta^2)^{-1/2} \tag{4.1}$$

in terms of the original parameterization. Even though Bernardo (1977) derived this prior for general sample sizes m and n, our results indicate that (4.1) requires $m = n$. We now investigate this problem outside the Normal context, retaining $\theta = \alpha/\beta$ as the parameter of interest and denoting the common scale parameter by σ (*i.e.* $\xi = \sigma$).

Theorem 2. *With the order $\{\theta, \beta, \sigma\}$ (three groups) and $\Theta^l_{\theta,\beta} = \Theta^l_\theta \times \Theta^l_\beta$ or $\Theta^l_{\theta,\beta}$ corresponding to $[-a(l), a(l)] \times [-b(l), b(l)]$ for (α, β), where $\{a(l)\}$ and $\{b(l)\}$ are increasing sequences of positive numbers, we obtain as the reference prior*

$$\pi(\alpha, \beta, \sigma) \propto \sigma^{-1} S^{-1/2}(\alpha, \beta), \tag{4.2}$$

where

$$\begin{aligned} S(\alpha, \beta) = {} & \left[\frac{m}{n} A(f)\left\{\frac{m}{n} c(f) + c(g)\right\} - \frac{m^2}{n^2} b^2(f)\right]\alpha^2 - 2\frac{m}{n} b(f) b(g) \alpha\beta \\ & + \left[A(g)\left\{\frac{m}{n} c(f) + c(g)\right\} - b^2(g)\right]\beta^2 \end{aligned} \tag{4.3}$$

with $A(\cdot)$, $b(\cdot)$ and $c(\cdot)$ as defined in $(B.1)$ taking $r = 1$.

The prior in (4.2) corresponds to $R(\alpha,\beta) = S^{-1/2}(\alpha,\beta)$ in Proposition 1 (i). The general expression for the reference prior is thus found to depend on the functional form of each sampling model as well as on the relative sample size. However, the expression for $S(\alpha,\beta)$ simplifies in the following situations:

Corollary 1. *If $b(f) = b(g) = 0$, the reference prior becomes*

$$\pi(\alpha,\beta,\sigma) \propto \sigma^{-1}\left(\frac{mA(f)}{nA(g)}\alpha^2 + \beta^2\right)^{-1/2}. \tag{4.4}$$

If, in addition, $f(\cdot) = g(\cdot)$, then we obtain

$$\pi(\alpha,\beta,\sigma) \propto \sigma^{-1}\left(\frac{m}{n}\alpha^2 + \beta^2\right)^{-1/2}. \tag{4.5}$$

In order to obtain the same reference prior for an entire class of sampling models, Corollary 1 imposes that $f(\cdot) = g(\cdot)$, apart from the condition $b(f) = 0$. Thus, for example, if both samples are generated from the same symmetric distribution, we have (4.5) as the reference prior. This, of course, applies to the original Fieller-Creasy problem, where Normality was assumed. The expression in (4.5) still depends on relative sample size, *i.e.* the definition of the underlying basic experiment. Only if $m = n$ do we obtain (4.1).

Again, the reference prior in (4.2) is not proper, which raises the issue of existence of the posterior. If we sample within the class of scale mixtures of Normals $b(f) = b(g) = 0$ and the prior in (4.2) – (4.3) simplifies to (4.4). The following result then holds.

Proposition 4. *Combining i.i.d. samples of $m \geq 1$ observations from (2.1) and $n \geq 1$ observations from (2.2), where both $f(\cdot)$ and $g(\cdot)$ correspond to scale mixtures of Normals and $\sigma = \xi$, with the prior (4.4) leads to a proper posterior if and only if $m + n \geq 3$. Furthermore, first and higher order posterior moments of $\theta = \alpha/\beta$ do not exist.*

5. THE PRODUCT OF LOCATIONS

We now consider a first sample of m univariate observations from (2.1) and a second sample of n univariate observations from (2.2) and focus our interest on the product of the locations. Thus, $\theta = \alpha\beta$ is the parameter of interest. Berger and Bernardo (1989) treat this problem with Normality imposed on both samples, with $\alpha, \beta \geq 0$ and known unitary variances ($\sigma = \xi = 1$). They comment on the difficulties encountered in classical estimation of $\theta = \alpha\beta$, when $\alpha, \beta \geq 0$. Implicitly, Berger and Bernardo (1989) assume both samples to be of equal size ($m = n$), whereas Bernardo and Smith (1994, Example 5.19) explicitly treat different sample sizes.

We shall extend their analysis in two stages. Firstly, we introduce general forms for $f(\cdot)$ and $g(\cdot)$, and secondly we consider unknown equal scales ($\sigma = \xi$). Overall, we shall choose as nuisance parameter $\rho = (\beta/\alpha)^{1/2}$. For the first extension, we assume $\sigma = \xi = 1$, but the analysis is trivially extended to any known values of σ and ξ as also mentioned in Berger and Bernardo (1989) for the Normal case. Our main results are summarized in the following theorems:

Theorem 3. Known scales
For m replications from (2.1) *and n replications from* (2.2) *with $\sigma = \xi = 1$, and two groups in the order $\{\theta, \rho\}$, the reference prior is:*
(i) *if $\Theta^l = \Theta^l_\theta \times \Theta^l_\rho$*

$$\pi(\alpha, \beta) \propto (\alpha\beta)^{-1} \left(\frac{m}{n} \frac{A(f)}{A(g)} \alpha^2 + \beta^2 \right)^{1/2}, \tag{5.1}$$

(ii) *if Θ^l corresponds to $[0, a(l)] \times [0, b(l)]$ for (α, β), where $\{a(l)\}$ and $\{b(l)\}$ are increasing sequences of positive numbers,*

$$\pi(\alpha, \beta) \propto \left(\frac{m}{n} \frac{A(f)}{A(g)} \alpha^2 + \beta^2 \right)^{1/2}, \tag{5.2}$$

with $A(\cdot)$ as defined in $(B.1)$ with $r = 1$.

As mentioned in Berger and Bernardo (1989) for the Normal case, the choice of the sequence of sets clearly matters. Both (i) and (ii) could be considered natural choices for Θ^l, but they lead to rather different forms for the reference prior. Generally, the choice of the sampling distributions also matters in (5.1) and (5.2), although it does not when $f(\cdot) = g(\cdot)$. Thus, the reference priors derived by Berger and Bernardo (1989) under Normality for $m = n$, extend to the case of taking both equal-sized samples from the same general location model. Berger and Bernardo (1989) intuitively favour taking rectangles in the original parameterization. Tibshirani (1989) builds on earlier results by Stein (1985) to give an asymptotic coverage probability motivation for the prior preferred by Berger and Bernardo (1989). In our more general framework (*i.e.* for any $f(\cdot)$ and $g(\cdot)$ and any sample sizes m and n), Tibshirani's method can also be applied leading to the prior in (5.2). Thus, from this point of view, the prior in (5.2) is preferable to the prior in (5.1).

There is still the issue of the existence of the posterior distribution left to resolve under both candidate reference priors. The following proposition collects results when both samples are taken from scale mixtures of Normals:

Proposition 5. *Using m observations from* (2.1) *and n observations from* (2.2), *where both $f(\cdot)$ and $g(\cdot)$ are scale mixtures of Normal density functions and $\sigma = \xi = 1$, the prior* (5.1) *will* **never** *lead to a proper posterior distribution. In contrast, the prior in* (5.2) *results in a proper posterior if* $\min\{m, n\} \geq 2$ *and posterior moments of θ of order $q \in [0, \min\{m, n\} - 2]$ are finite.*

Thus, under the prior in (5.1) posterior inference is precluded, for any sample sizes and sampling densities within the entire class of scale mixtures of Normals. This immediately implies that the posterior distribution does not exist under Normal sampling either, a result that is implicit in Sun and Ye (1995), who consider a product of many Normal means. On the basis of plots of the (nonexisting) posterior, Berger and Bernardo (1989) conclude that the prior in (5.1) is "highly counterintuitive". We could thus considerably strengthen this to stating that (5.1) can not lead to posterior inference. Note that this fully resolves the choice between (5.1) and (5.2) (at least under a wide and relevant class of sampling distributions) and the question in which parameterization one should choose rectangles becomes moot in this case.

Theorem 4. Unknown equal scales
We consider samples of sizes m and n from (2.1) and (2.2), respectively, with $\sigma = \xi$ unknown and the three groups in the order $\{\theta, \rho, \sigma\}$. If either or both of $b(f)$ and $b(g)$ is zero, the reference prior is:
(i) *if* $\Theta^l_{\theta,\rho} = \Theta^l_\theta \times \Theta^l_\rho$

$$\pi(\alpha, \beta, \sigma) \propto \sigma^{-1}(\alpha\beta)^{-1} \left(\frac{m}{n} \frac{M(f, g, \frac{n}{m})}{M(g, f, \frac{m}{n})} \alpha^2 + \beta^2 \right)^{1/2}, \tag{5.3}$$

(ii) *if* $\Theta^l_{\theta,\rho}$ *corresponds to* $[0, a(l)] \times [0, b(l)]$ *for* (α, β),

$$\pi(\alpha, \beta, \sigma) \propto \sigma^{-1} \left(\frac{m}{n} \frac{M(f, g, \frac{n}{m})}{M(g, f, \frac{m}{n})} \alpha^2 + \beta^2 \right)^{1/2}, \tag{5.4}$$

where

$$M(f, g, \frac{n}{m}) = A(f) - \frac{b^2(f)}{c(f) + \frac{n}{m} c(g)}, \tag{5.5}$$

with $A(\cdot)$, $b(\cdot)$ *and* $c(\cdot)$ *as in* $(B.1)$ *with* $r = 1$.

Note that if $b(f) = 0$, then $M(f, g, n/m) = A(f)$. Thus, if both $b(f)$ and $b(g)$ are zero, we obtain the same expressions as in Theorem 3 with an additional factor σ^{-1}. Just like in Theorem 3, the choice of $\{\Theta^l\}$ influences the reference prior. This contrasts with Theorem 1 for the difference and Theorem 2 for the ratio of locations, where both choices of $\{\Theta^l\}$ led to the same form of the reference prior. The problem of products of locations is particularly sensitive to the choice of the sequence of sets, as already noted by Berger and Bernardo (1989) in the Normal context. Theorem 4 presents another example of Proposition 1 (i), where the form of the function $R(\cdot)$ now also depends on the sequence $\{\Theta^l_{\theta,\rho}\}$.

Let us now check for existence of the posterior distribution and moments in this case with equal unknown scales.

Proposition 6. *Combining m observations from* (2.1) *and n observations from* (2.2), *where both $f(\cdot)$ and $g(\cdot)$ are scale mixtures of Normal density functions and $\sigma = \xi$, with the prior* (5.3) *will* **never** *lead to a proper posterior distribution. If we use the prior in* (5.4) *we obtain a proper posterior if* $\min\{m, n\} \geq 2$ *and posterior moments of θ of order* $q \in [0, (\min\{m, n\} - 2)/2]$ *are finite.*

Just like in the case of known scales, taking the sets $\Theta^l_{\theta,\rho} = \Theta^l_\theta \times \Theta^l_\rho$ leads to a nonexistent posterior distribution for a large and practically useful class of sampling distributions. Thus, the choice between (5.3) and (5.4) can immediately be decided, at least when sampling from scale mixtures of Normals.

As a final remark, the conditions stated in Propositions 5 and 6 under the priors (5.2) and (5.4) are merely sufficient conditions that need not be necessary in general.

6. EMPIRICAL ILLUSTRATION

In this section we shall use the reference priors derived previously in analyzing a set of data collected in a clinical trial. Karpatkin, Porges and Karpatkin (1981) investigate the impact of maternal steroid therapy on the platelet counts of newborn infants. The data are here taken in thousands of platelets per mm^3, and are also reported in Pocock (1983). We have univariate observations ($r = s = 1$) on the infant platelet count after delivery for $m = 7$ mothers not given steroids and $n = 12$ mothers that were given steroids. In this application, it seems natural to assume independence between both samples. Furthermore, we shall assume the samples were i.i.d. drawings from the location-scale models in (2.1) and (2.2), respectively, and we will now focus on the difference and the ratio of the locations. The data for mothers without steroids, $(x_1, \dots, x_m)$, display some skewness (Pearson's measure of skewness, γ_1, is 1.00), whereas the second sample, $(y_1, \dots, y_n)$, has even more skewness ($\gamma_1 = 1.53$) and also possesses excess kurtosis ($\gamma_2 = 1.74$). Hence, a rank-based procedure was proposed in Pocock (1983).

6.1. The Difference of Locations

Often, this problem is posed under Normality with unknown and possibly different variances (Behrens-Fisher problem). Here we shall also investigate the problem under alternative assumptions for $f(\cdot)$ and $g(\cdot)$: Student-t distributions, which give thicker tails, and Skewed Exponential Power distributions with $q = 2$, thus introducing skewness [see $(B.2)$ and $(B.4)$ in Appendix B]. Since these sampling distributions all lead to block-diagonal information matrices [given in $(B.3)$ and $(B.5)$, respectively], Proposition 2 applies and the reference prior is (3.1) for all models. Due to the product structure of (3.1) and the independence of both samples, we know that (α, σ) and (β, ξ) (and thus α and β) are a posteriori independent. Thus, it suffices to present the derivation of the posterior distribution for just one sample. Then, different stochastic assumptions for x in (2.1) and y in (2.2) can immediately be combined. Let us thus only describe the analysis of the posterior distribution of α, computed under m replications from (2.1) with the prior $\pi(\alpha, \sigma) \propto \sigma^{-1}$. From independent drawings on α and β we directly construct drawings for $\theta = \alpha - \beta$, which is the focus of interest. Throughout this subsection, our empirical results will be based on 50,000 drawings. In cases (ii) and (iii) drawings on α are generated using a Gibbs sampler on (α, σ), possibly augmented with some other variables.

(i) *NORMAL SAMPLING*

It is well-known that if $f(\cdot)$ is a Normal p.d.f., the marginal posterior distribution of α is a Student-t distribution with $m - 1$ degrees of freedom. Clearly, moments will exist up to the order $m - 1$ (not including).

(ii) *STUDENT SAMPLING WITH ν DEGREES OF FREEDOM*

This corresponds to sampling from (2.1) with $f(\cdot)$ as given in $(B.2)$ (choosing $r = 1$). Contrary to the previous case, a tractable analytical expression for the marginal posterior density of α does not exist, and we shall resort to Gibbs sampling (see Gelfand and Smith, 1990; Casella and George, 1992). From Proposition 3, the posterior density of α is proper if and only if $m \geq 2$ and has moments up to and not including $m-1$. Exploiting the representation of a Student-t distribution as a scale mixture of Normals [with P_λ in (3.2) a Gamma$(\nu/2, \nu/2)$ distribution], a Gibbs sampler on α, σ and the mixing variables $\lambda_1, \dots, \lambda_m$ can easily be set up as described in *e.g.* Geweke (1993).

(iii) *SKEWED EXPONENTIAL POWER SAMPLING*

Finally, we introduce skewness into the sampling distribution by considering $f(\cdot)$ given in $(B.4)$ for $r = 1$ and $q = 2$ (*i.e.* a "Skewed Normal" as defined in Fernández and Steel, 1998). The marginal posterior density of α is then a mixture of truncated Student-t distributions with $m - 1$ degrees of freedom. Therefore, it is proper provided sample size $m \geq 2$, and has moments up to (and not including) $m - 1$, just as in both previous cases. In addition, we can prove that its density, $\pi(\alpha|x_1, \ldots, x_m)$, is continuous and first order differentiable, despite being a mixture of truncated densities. In order to generate drawings from this non-standard distribution for α, we shall use a Gibbs sampler on α and σ, which is a simpler version of the one devised by Fernández and Steel (1998) for Skewed Student sampling.

Figure 1 plots the posterior density function of $\theta = \alpha - \beta$ on the basis of the platelet data and for three different sampling schemes. First, we consider Normality for both samples, *i.e.* the usual Behrens-Fisher problem. Then, we introduce thicker tails and, in view of the data kurtosis in both samples, we choose to use a Student-t with ten degrees of freedom for $f(\cdot)$ and a Cauchy for $g(\cdot)$. This change of sampling distributions has a rather marked effect on the posterior of θ. Finally, we acknowledge the fact that the data display considerable positive skewness, as discussed above. In order to account for this, we use the Skewed Exponential Power (SEP) distribution in $(B.4)$ with $q = 2$, and take $\gamma = 1.5$ for $f(\cdot)$ and $\gamma = 2$ for $g(\cdot)$. We then obtain the third posterior density of θ displayed in Figure 1. From the previous theory, we know that all three posterior distributions have moments up to $\min\{m - 1, n - 1\}$, which is six in our case.

Apart from the difference of the locations (which correspond to the modes under all three sampling schemes), we might be interested in the mean difference between the counts in both samples. For symmetric choices of $f(\cdot)$ and $g(\cdot)$ these are equivalent (provided the mean exists), but for the SEP this is not the case. Figure 1 also indicates the mean difference under the SEP specifications, and, as expected, it is much closer to the Normal result than the difference between the modes. Of course, if our interest would really be the mean difference, rather than the difference between the modes, we would have to use the appropriate reference prior for that parameter of interest.

For illustrative purposes we have fixed the values of the degrees of freedom, ν, in the Student case, and the skewness parameter, γ, for the SEP specification, roughly in accordance with the characteristics of the data. In serious applications, we would, of course, recommend estimating these parameters jointly with location and scale parameters, as described in Fernández and Steel (1998) in the context of Skewed Student-t sampling. Note, however, that treating γ and ν as unknown parameters to be estimated, could change the form of the reference prior for the location and scale parameters.

Some sensitivity with respect to sampling assumptions is directly apparent from Figure 1. While all sampling schemes would agree to a positive effect of steroids on the modal platelet count, they vary quite a bit as to the size of such an effect.

6.2. The Ratio of Locations

In this subsection we shall first treat the usual Fieller-Creasy problem with Normal sampling and unknown equal variances. In addition, we shall examine the case where the first sample is still assumed to come from a Normal distribution, whereas for the second sample we take a Cauchy distribution. Note that here the common scale, denoted by σ ($\sigma = \xi$), links both samples and thus we need to consider the joint posterior of α and β, from which drawings of $\theta = \alpha/\beta$ will be generated. By changing $g(\cdot)$ to a Cauchy distribution (for which even the mean does not exist) the assumption of a common scale parameter is no longer in flagrant contradiction with the fact that the variance of the second sample is much larger than that of the first.

Applying Corollary 1, the reference prior is given by (4.5) under Normality for both samples, and by

$$\pi(\alpha, \beta, \sigma) \propto \sigma^{-1} \left(\frac{m}{n} 2\alpha^2 + \beta^2 \right)^{-1/2}, \tag{6.1}$$

when $g(\cdot)$ is replaced by a Cauchy p.d.f. Since the posterior analysis greatly simplifies if we use instead of (4.5) or (6.1) the independence Jeffreys' prior

$$\pi(\alpha, \beta, \sigma) \propto \sigma^{-1}, \tag{6.2}$$

we adopt the following strategy outlined in Stephens and Smith (1992):

First, we use (6.2) and generate drawings from the corresponding marginal posterior distribution for (α, β). In a second stage we resample (with replacement) from the set of generated drawings with weights proportional to $\{(m/n)\alpha^2 + \beta^2\}^{-1/2}$ for the Normal case, and $\{(m/n)2\alpha^2 + \beta^2\}^{-1/2}$ for the Normal-Cauchy sampling. This Sampling-Importance Resampling (SIR) technique will then generate drawings from the posterior distributions of (α, β) under the priors (4.5) and (6.1). In the actual computations we used a sample of 150,000 values from which we resampled 10,000 drawings of (α, β), which were then simply transformed to drawings of $\theta = \alpha/\beta$.

(i) *NORMAL SAMPLING*

Under the simple prior in (6.2), it is easy to derive that the posterior distribution $\pi(\alpha, \beta | data)$ is a bivariate Student with $m + n - 2$ degrees of freedom. The numerical analysis will now be conducted by drawing α and β from this bivariate Student-t distribution and resampling with the weights corresponding to the second factor in (4.5). We note that this SIR scheme requires propriety of the Student-t distribution, which translates to $m + n \geq 3$, *i.e.* a total of at least three observations. This is exactly the necessary and sufficient condition for a proper posterior distribution, as explained in Proposition 4.

(ii) *NORMAL-CAUCHY SAMPLING*

We now assume that the second sample $(y_1, \ldots, y_n)$ is generated from a Cauchy distribution. Thus, following (3.2),

$$p(y_i | \beta, \sigma, \lambda_i) = f_N^1(y_i | \beta, \sigma^2 \lambda_i^{-1}), \quad i = 1, \ldots, n,$$

where each λ_i has an independent Gamma(1/2,1/2) distribution and $f_N^r(z|a, B)$ corresponds to an r-variate Normal distribution with mean a and covariance matrix B.

Proposition 4 establishes propriety of the posterior distribution under the prior in (6.1) if and only if $m + n \geq 3$. The same can be shown to hold under the prior (6.2). Now, even with the latter prior, the posterior distribution of (α, β) displays a non-standard form, requiring numerical techniques for its analysis. We shall use a Gibbs sampler on $(\alpha, \beta, \sigma, \lambda_1, \ldots, \lambda_n)$ through the following conditionals [under (6.2)]:

$$\pi\left(\alpha, \beta, \sigma^{-2} | \lambda_1, \ldots, \lambda_n, data\right) = f_N^2 \left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix} | \begin{pmatrix} \sum_{i=1}^m x_i/m \\ \frac{\sum_{i=1}^n \lambda_i y_i}{\sum_{i=1}^n \lambda_i} \end{pmatrix}, \sigma^2 \begin{pmatrix} m & 0 \\ 0 & \sum_{i=1}^n \lambda_i \end{pmatrix}^{-1} \right)$$
$$\times f_G \left(\sigma^{-2} | \frac{m+n-2}{2}, \frac{\sum_{i<j}(x_i - x_j)^2}{2m} + \frac{\sum_{i<j} \lambda_i \lambda_j (y_i - y_j)^2}{2 \sum_{i=1}^n \lambda_i} \right), \tag{6.3}$$

where $f_G(w|c,d)$ is the p.d.f. of a Gamma distribution with shape parameter c and mean c/d, and

$$\pi(\lambda_1, \ldots, \lambda_n | \alpha, \beta, \sigma^{-2}, data) = \prod_{i=1}^n f_G \left(\lambda_i | 1, \frac{1 + \sigma^{-2}(y_i - \beta)^2}{2} \right). \tag{6.4}$$

This Gibbs sampler is used to generate a set of drawings from (α, β), which is then resampled with weights proportional to $\{(m/n)2\alpha^2 + \beta^2\}^{-1/2}$ in order to obtain drawings from the marginal posterior distribution of (α, β) with the reference prior in (6.1).

Figure 2 summarizes the posterior inference on $\theta = \alpha/\beta$ using the infant platelet data described above. We note that the Normal-Cauchy sampling leads to quite different posterior inference on θ than the usual Normal assumption. From Proposition 4, neither of these distributions possesses a finite first order moment. Both sampling specifications clearly indicate that the use of steroids increases platelet counts, since $pr(\theta < 1|data)$ is virtually equal to unity. In order to investigate whether the use of steroids doubles the platelet count, we can compute:

$$pr\left(\theta < \frac{1}{2} | data\right) = pr(\beta > 2\alpha | data) = \begin{cases} 0.999 & \text{for Normal sampling} \\ 0.874 & \text{for Normal-Cauchy sampling.} \end{cases}$$

7. CONCLUSIONS

This paper investigates the reference prior in the context of general multivariate continuous location-scale models, with unknown location and scale. In particular, we consider the situation where two samples were obtained independently from different location-scale models. The forms of the sampling densities are kept entirely free, subject to regularity conditions assuring asymptotic Normality of the likelihood functions. Both samples can be of different sizes and generated from different sampling distributions with p.d.f.'s $f(\cdot)$ and $g(\cdot)$. The parameter of interest, denoted by θ, will be chosen as some function of both locations, α and β. We treat in detail the difference, ratio and product of locations, and thus generalize, respectively, the Behrens-Fisher, Fieller-Creasy and product of Normal means problems. Generally, we show that the forms of the sampling distributions, the ratio of the sample sizes, and the choices of θ and the sequence of sets $\{\Theta^l\}$ can influence the reference prior.

In contrast to the influence of the forms of the sampling distributions and the choice of the parameter of interest, the fact that the relative sample size m/n can affect the reference prior is perhaps somewhat counterintuitive. From a purely formal point of view, this arises through the definition of the basic experiment and its information matrix. Intuitively, it is not surprising that the way one accumulates sample information about the parameter of interest may depend on m/n, and thus a prior that is expressly designed to maximize this information may well posses the same type of dependence. For the difference of locations, relative sample size is found not to influence the reference prior, but it always intervenes in the other two problems.

A contentious issue in the application of the reference prior algorithm is the choice of the sequence of nested compact sets $\{\Theta^l\}$. It is often felt reasonable to choose a Cartesian product for Θ^l (such as rectangles), but then it is not always clear whether rectangles in the original parameters (locations) or in the transformed parameters (including the parameter of interest) are preferable. For the difference and ratio of locations, both sequences lead to exactly the same reference prior. If we are interested in the product of locations, however, the same invariance no longer holds. Berger and Bernardo (1989) address this situation in the context of a product of Normal means. We extend their results to general sampling distributions and unequal sample sizes. The reference prior corresponding to rectangles in the locations (*i.e.* original parameterization), can be given an asymptotic coverage probability motivation, extending the analysis in Tibshirani (1989). An even more compelling argument in favour of the latter prior is the fact that for a wide class of sampling distributions, namely scale mixtures of Normals, the prior corresponding to rectangles in the transformed parameters does not lead to a proper posterior distribution. Thus, for many practically interesting sampling models this prior could never lead to posterior inference. The same fundamental problem occurs if we allow for an unknown scale factor common to both sampling models. Therefore, the apparent need to make an essentially arbitrary choice for the reference prior on the basis of Θ^l vanishes in the case of all two-sample problems considered here.

In addition to the influences that are formally examined in this paper, there are several other issues that can potentially affect the form of the reference prior. The algorithm also depends on the number of parameter groups and their ordering (apart from the fact that we take the parameter of interest as the first group), and here we have generally followed the strategy of putting each parameter in a separate group (as advocated by Berger and Bernardo, 1992) and we have mostly ordered the scales last. In addition, whereas the choice of the nuisance parameter is irrelevant when there are only two groups (see Proposition 5.27 of Bernardo and Smith, 1994), we know of no general results in this respect when the number of groups is larger than two (as is mostly the case in this paper).

We use medical data on the effect of maternal steroid treatment on the platelet count of newborn infants as an illustration of the generalized Behrens-Fisher and Fieller-Creasy problems. For the difference of locations we contrast Normal sampling with independent sampling from Student and Skewed Exponential Power distributions. All three cases share a common form of the reference prior. Both departures from Normality clearly affect the posterior inference on $\theta = \alpha - \beta$. For the case where $\theta = \alpha/\beta$, we first take both sampling distributions to be Normal and we then consider the situation where $f(\cdot)$ is Normal and

$g(\cdot)$ is a Cauchy p.d.f. The reference prior now varies with the sampling distribution. Again, the posterior distribution of θ is substantially affected by the choice of sampling scheme.

Our empirical exercise illustrates both the feasibility of posterior analysis under non-Normal sampling distributions, using recently developed numerical techniques, and the potential sensitivity of posterior inference to changes in the sampling distributions. Thus, extending the reference posterior analysis to non-Normal sampling distributions is of genuine practical interest. We feel our theoretical results on the influence of the sampling scheme and the choices of $\{\Theta^l\}$ and θ on the reference prior also increase our understanding of the intricate workings of this algorithm.

APPENDIX A: SKETCHED PROOFS

Proposition 1: We present the proof for $\sigma = \xi$ (the proof for the case of different scales can be done along the same lines). The information matrix for (α, β, σ) takes the form

$$I(\alpha, \beta, \sigma) = \sigma^{-2} n B(f, g, m/n), \qquad (A.1)$$

for some $(r+s+1) \times (r+s+1)$ matrix $B(\cdot)$. Using $(A.1)$ and the suitable Jacobian, we obtain the information matrix for (θ, ρ, σ). We now apply the reference prior algorithm of Berger and Bernardo (1992) following their notation:

Since $h_3 \propto \sigma^{-2}$, we obtain $\pi_3^l(\sigma|\theta, \rho) \propto \sigma^{-1}$. The marginal prior for (θ, ρ) obviously does not depend on σ, from which Proposition 1 (i) follows directly.

Theorem 1: Applying the reference prior algorithm to the information matrix of $(\theta, \beta, \sigma, \xi)$ following the notation of Berger and Bernardo (1992), we obtain $h_4 \propto \xi^{-2}$, $h_3 \propto \sigma^{-2}$, whereas h_2 and h_1 do not depend on θ or β. This implies for either choice of Θ^l in Theorem 1

$$\pi_1^l(\theta, \beta, \sigma, \xi) \propto \frac{\sigma^{-1}\xi^{-1}}{\int_{\Theta^l(\theta)} d\beta}, \text{ where } \Theta^l(\theta) = \{\beta : (\theta, \beta) \in \Theta^l_{\theta,\beta}\}. \qquad (A.2)$$

When $\Theta^l_{\theta,\beta} = \Theta^l_\theta \times \Theta^l_\beta$, the integral in $(A.2)$ is constant, thus obtaining (3.1) as the reference prior. For the second choice of $\Theta^l_{\theta,\beta}$ in Theorem 1, the integral in $(A.2)$ can depend on θ, but as $\int_{\Theta^l(\theta=0)} d\beta / \int_{\Theta^l(\theta)} d\beta$ converges to one as $l \to \infty$, we again obtain (3.1) as the reference prior. The proof for the order $\{\theta, \beta, \xi, \sigma\}$ can be done in a similar way.

Proposition 2: Similar to the proof of Theorem 1.

Proposition 3: Consider the product of $|\alpha_j|^q$ and the joint distribution of the observables, α, σ and the mixing parameters $\lambda_1, \ldots, \lambda_m$ from (3.2). After integrating out α and σ, which requires $q < r(m-1)$, we are left with a bounded (for almost any sample) function of $(\lambda_1, \ldots, \lambda_m)$, which is therefore integrable with respect to any mixing distribution.

Theorem 2: Applying the reference prior algorithm on the information matrix of (θ, β, σ), we obtain $h_3 \propto \sigma^{-2}$, whereas h_2 does not depend on β, and $h_1 \propto \beta^2 \sigma^{-2} S^{-1}(\theta, 1)$, with $S(\cdot)$ defined in (4.3). This leads to

$$\pi_1^l(\theta, \beta, \sigma) \propto \frac{\sigma^{-1} S^{-1/2}(\theta, 1)}{\int_{\Theta^l(\theta)} d\beta} \exp\left\{ \frac{\int_{\Theta^l(\theta)} \log \beta^2 \, d\beta}{2 \int_{\Theta^l(\theta)} d\beta} \right\}, \qquad (A.3)$$

with $\Theta^l(\theta)$ defined in $(A.2)$, for either choice of Θ^l in Theorem 2. When $\Theta^l_{\theta,\beta} = \Theta^l_\theta \times \Theta^l_\beta$, each of the three integrals in $(A.3)$ is a constant, thus obtaining the reference prior in (4.2) -after a transformation to the original parameterization-. Under the second choice of $\Theta^l_{\theta,\beta}$ in Theorem 2, each of the integrals in $(A.3)$ separately depends on θ but this dependence cancels between them, so that the expression in $(A.3)$ is still proportional to $\sigma^{-1}S^{-1/2}(\theta, 1)$.

Proposition 4: Consider the joint distribution of the observables, the mixing parameters [see (3.2)], and (θ, β, σ). We first integrate out σ^{-2} as a Gamma, then β as a Student-t, leaving us with a posterior density for θ given the scale mixing parameters from both samples which has a lower bound proportional to a Cauchy p.d.f. This explains the nonexistence of posterior moments of order one and higher. Finally, propriety for $m = n = 1$ never holds, whereas it can be established for $m+n = 3$ with any scale mixing distributions (for almost any sample).

Theorem 3: Direct application of the reference prior algorithm starting from the information matrix of (θ, ρ), leads to

$$\pi^l_1(\theta,\rho) \propto \frac{\theta^{-1/2}(A\rho^{-4}+1)^{1/2}}{\int_{\Theta^l(\theta)}(A\rho^{-4}+1)^{1/2}d\rho} \exp\left[-\frac{\int_{\Theta^l(\theta)}(A\rho^{-4}+1)^{1/2}\log(A\rho^{-2}+\rho^2)d\rho}{2\int_{\Theta^l(\theta)}(A\rho^{-4}+1)^{1/2}d\rho}\right], \quad (A.4)$$

where $\Theta^l(\theta) = \{\rho : (\theta, \rho) \in \Theta^l\}$ and $A = \{mA(f)\}/\{nA(g)\}$ with $A(f)$ and $A(g)$ defined through $(B.1)$. Theorem 3 (i) is immediate from $(A.4)$. In order to prove Theorem 3 (ii), we make the change of variables $\omega = \rho A^{-1/4}$ in each of the three integrals in $(A.4)$. The resulting expression in $(A.4)$ is then very close to that obtained under Normality, which directly allows us to use the results in Section 5.2 of Berger and Bernardo (1989) to prove Theorem 3 (ii).

Proposition 5: We consider the joint distribution of the observables, the mixing parameters and (α, β). The prior (5.1) has a lower bound proportional to $\max\{1/\alpha, 1/\beta\}$ and thus would require first order negative moments of a Normal to exist, which directly leads to the result. For the prior (5.2) we use an upper bound proportional to $\max\{\alpha, \beta\}$. After integrating out α and β, we are then left with a function of the mixing parameters, which is bounded (for almost any sample) provided that $q \leq \min\{m, n\} - 2$.

Theorem 4: After a first step that leads to the factor σ^{-1}, we follow the proof of Theorem 3, replacing A in $(A.4)$ by $\{mM(f, g, n/m)\}/\{nM(g, f, m/n)\}$ with $M(\cdot)$ defined in (5.5).

Proposition 6: Similar to the proof of Proposition 5, using a lower bound proportional to $\sigma^{-1}\max\{1/\alpha, 1/\beta\}$ for (5.3) and an upper bound proportional to $\sigma^{-1}\max\{\alpha, \beta\}$ for (5.4).

APPENDIX B: SOME INFORMATION MATRICES

General location-scale model

The information matrix for the general location-scale model in (2.1) takes the form

$$I(\alpha, \sigma) = \sigma^{-2}\begin{pmatrix} A(f) & b(f) \\ b'(f) & c(f) \end{pmatrix}, \quad (B.1)$$

where $A(f)$ is an $r \times r$ matrix, $b(f) \in \Re^r$ and $c(f) > 0$. If $f(\cdot)$ is axially symmetric, we can derive that $I(\alpha, \sigma)$ is diagonal. If, in addition, $f(\cdot)$ possesses exchangeability [*i.e.* $f(z_1, \ldots, z_r) = f(z_{\phi(1)}, \ldots, z_{\phi(r)})$ for any permutation $\{\phi(1), \ldots, \phi(r)\}$ of $\{1, \ldots, r\}$], then $A(f)$ is a multiple of I_r, the identity matrix of rank r.

Student-t distribution

If, in the spherical context, we consider the r-variate Student-t with $\nu > 0$ degrees of freedom, which corresponds to

$$f(z) = \frac{\Gamma\left(\frac{\nu+r}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{r/2}} \left(1 + \frac{1}{\nu}\|z\|^2\right)^{-\frac{\nu+r}{2}}, \tag{$B.2$}$$

the following information matrix results in $(B.1)$:

$$I(\alpha, \sigma) = \sigma^{-2} \begin{pmatrix} \frac{\nu+r}{\nu+r+2} I_r & 0 \\ 0 & 2r\frac{\nu}{\nu+r+2} \end{pmatrix}. \tag{$B.3$}$$

Note that, as $\nu \to \infty$, the information matrix in $(B.3)$ converges to that of the Normal distribution.

Skewed Exponential Power distribution

This family of distributions was introduced by Fernández, Osiewalski and Steel (1995), and in an r-variate context corresponds to

$$f(z) = \left\{ \frac{q}{2^{1/q}\Gamma\left(\frac{1}{q}\right)\left(\gamma + \frac{1}{\gamma}\right)} \right\}^r \exp\left[-\frac{1}{2} \sum_{i=1}^{r} \left\{ \left(\frac{z_i}{\gamma}\right)^q I_{[0,\infty)}(z_i) + (-\gamma z_i)^q I_{(-\infty,0)}(z_i) \right\} \right], \tag{$B.4$}$$

where $\gamma \in (0, \infty)$. For $\gamma = 1$, $(B.4)$ corresponds to r i.i.d. replications from a univariate Exponential Power distribution, and $f(\cdot)$ is therefore exchangeable and axially symmetric. However, for $\gamma \neq 1$, $f(\cdot)$ no longer possesses axial symmetry. It can be shown that the value of γ in $(B.4)$ does not affect the information matrix in $(B.1)$, which takes the form

$$I(\alpha, \sigma) = \sigma^{-2} \begin{pmatrix} \frac{q(q-1)}{4^{1/q}} \frac{\Gamma\left(1-\frac{1}{q}\right)}{\Gamma\left(\frac{1}{q}\right)} I_r & 0 \\ 0 & rq \end{pmatrix}. \tag{$B.5$}$$

Clearly, for $q = 2$, we recuperate the information matrix of the Normal case.

REFERENCES

Berger, J.O., and Bernardo, J.M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84,** 200-207.

Berger, J.O., and Bernardo, J.M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.). Oxford: Oxford University Press, 35-60 (with discussion).

Bernardo, J.M. (1977). Inferences about the ratio of Normal means: A Bayesian approach to the Fieller-Creasy problem. *Recent Developments in Statistics* (J.R. Barra,

F. Brodeau, G. Romier, and B. Van Cutsem, eds.). Amsterdam: North-Holland, 345-350.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113-147 (with discussion).

Bernardo, J.M., and Smith, A.F.M. (1994). *Bayesian Theory.* Chichester: John Wiley.

Casella, G., and George, E. (1992). Explaining the Gibbs sampler. *Am. Statistician* **46,** 167-174.

Clarke, B. (1996). Implications of reference priors for prior information and for sample size. *J. Amer. Statist. Assoc.* **91**, 173-184.

Clarke, B., and Wasserman, L. (1993). Noninformative priors and nuisance parameters. *J. Amer. Statist. Assoc.* **88**, 1427-1432.

Creasy, M.A. (1954). Limits for the ratio of the means (with discussion). *J. Roy. Statist. Soc. B* **16**, 186-194.

Datta, G.S., and Ghosh, M. (1995). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357-1363.

DeGroot, M.H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

Fernández, C., Osiewalski, J., and Steel, M.F.J. (1995). Modeling and inference with v-spherical distributions. *J. Amer. Statist. Assoc.* **90**, 1331-1340.

Fernández, C., and Steel, M.F.J. (1998). On Bayesian modeling of fat tails and skewness. *J. Amer. Statist. Assoc.* **93**, forthcoming.

Fieller, E.C. (1954). Some problems in interval estimation. *J. Roy. Statist. Soc. B* **16**, 175-185 (with discussion).

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd.

Florens, J.P., Mouchart, M. and Rolin, J.M. (1990). Invariance arguments in Bayesian statistics. *Economic Decision Making: Games, Econometrics and Optimisation* (J. Gabsewicz, J.F. Richard and L.A. Wolsey, eds.). Amsterdam: North-Holland, 351-367.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *J. Appl. Econometrics* **8,** S19-S40.

Ghosal, S. (1997). Reference priors in multiparameter nonregular cases. *Test* **6**, 159-186.

Ghosh, J.K., and Mukerjee, R. (1992). Non-informative priors. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.) Oxford: Oxford University Press, 321-344 (with discussion).

Jeffreys, H. (1961). *Theory of Probability.* Oxford: Oxford University Press, third edition.

Kass, R.E., and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343-1370.

Karpatkin, M., Porges, R.F., and Karpatkin, S. (1981). Platelet counts in infants of women with autoimmune thrombocytopenia, Effect of steroid administration to the mother. *New Engl. J. Med.* **305**, 936-937.

Liseo, B. (1992). L'uso della funzione di verosimiglianza negli approcci condizionati all'inferenza statistica. Ph.D. Thesis, Dipartimento di Statistica, Probabilita e Statistiche Applicate, Universita di Roma.

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295-304.

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

Pocock, S.J. (1983). *Clinical Trials: A Practical Approach.* Chicester: John Wiley.

Stein, C. (1985). On the coverage probability of confidence sets based on a prior distribution. *Sequential Methods in Statistics* (R. Zielisni, ed.). Warsaw: PWN-Polish scientific publishers, 485-514.

Stephens, D.A., and Smith, A.F.M. (1992). Sampling-resampling techniques for the computation of posterior densities in Normal means problems. *Test* **1**, 1-18.

Sun, D., and Ye, K. (1995). Reference prior Bayesian analysis for Normal means products. *J. Amer. Statist. Assoc.* **90**, 589-597.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76,** 604-608.

Figure 1: Posterior density of difference of locations

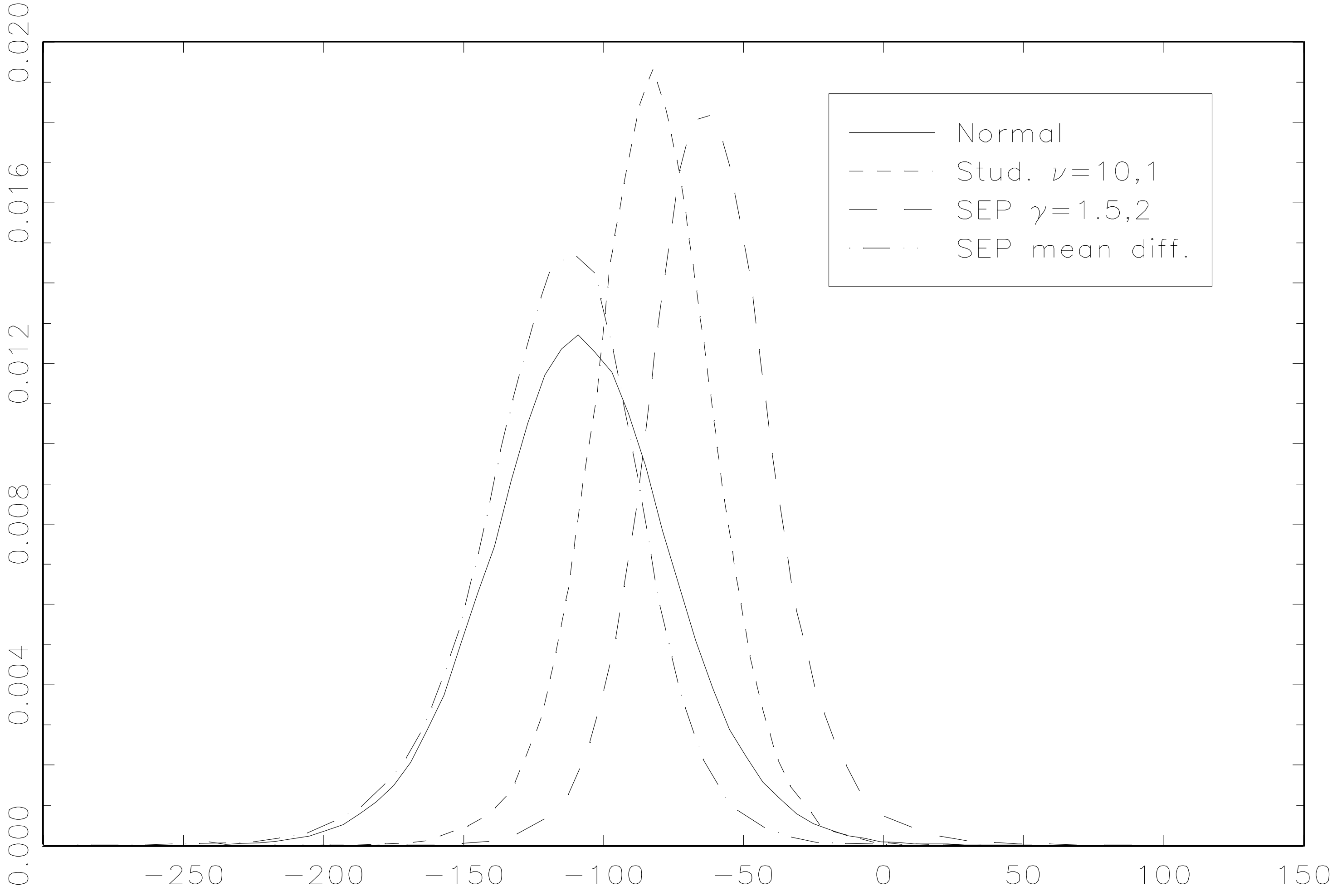
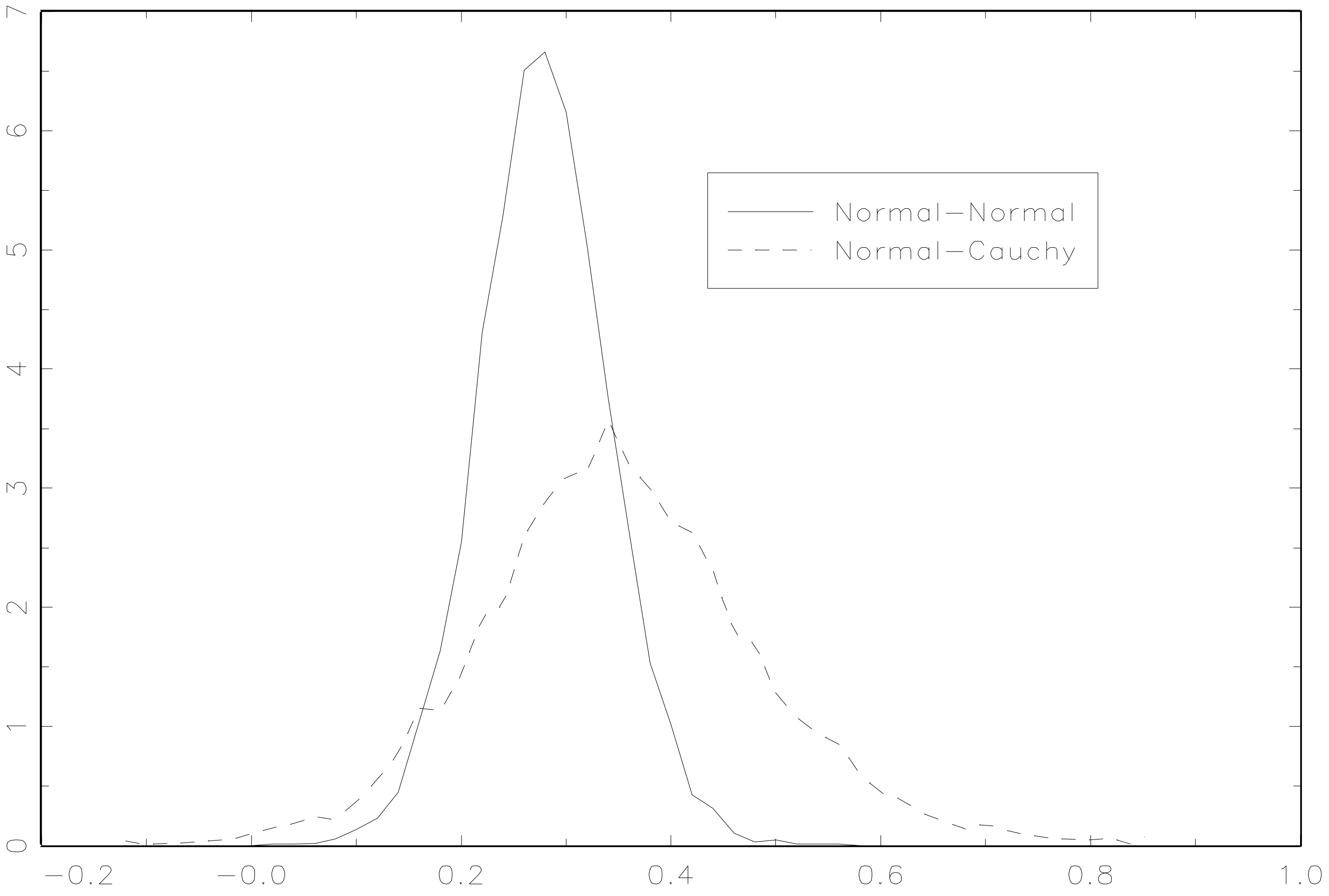


Figure 2: Posterior density of ratio of locations


Normal−Normal
Normal−Cauchy