BAYESIAN STATISTICS 6, pp. 1–16
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)

On the Dangers of Modelling Through Continuous Distributions: A Bayesian Perspective

CARMEN FERNANDEZ and MARK F. J. STEEL
University of Bristol, U.K. and
University of Edinburgh, U.K.

SUMMARY

We point out that Bayesian inference on the basis of a given sample is not always possible with continuous sampling models, even under a proper prior. The reason for this paradoxical situation is explained, and linked to the fact that any dataset consisting of point observations has zero probability under a continuous sampling distribution. A number of examples, both with proper and improper priors, highlight the issues involved. A solution is proposed through the use of set observations, which take into account the precision with which the data were recorded. Use of a Gibbs sampler makes the solution practically feasible. The case of independent sampling from (possibly skewed) scale mixtures of Normals is analysed in detail for a location-scale model with a commonly used noninformative prior. For Student-t sampling with unrestricted degrees of freedom the usual inference, based on point observations, is shown to be precluded whenever the sample contains repeated observations. We show that Bayesian inference based on set observations, however, is possible and illustrate this by an application to a skewed dataset of stock returns.

Keywords: LOCATION-SCALE MODEL; ROUNDING; SCALE MIXTURES OF NORMALS; SKEWNESS; STUDENT-T.

1. INTRODUCTION

The purpose of this paper is to examine some pathological situations that may arise when we conduct Bayesian inference using a continuous sampling distribution (*i.e.* absolutely continuous with respect to Lebesgue measure). As examples of the rather counterintuitive phenomena that can occur, let us mention that even under a proper prior, posterior inference could be precluded for certain samples, and that adding new observations could destroy the possibility of conducting inference while it was feasible with the previous sample. Whereas under a proper prior problems can only arise if the likelihood function is unbounded, they can be encountered both with bounded and unbounded likelihoods if an improper prior is used.

Since the Bayesian paradigm solely relies on probability and measure theory, we shall examine the pitfalls mentioned above within this framework. The first thing to notice is that a sample of point observations (which is the way in which most datasets are recorded) has zero probability under any continuous sampling distribution. Such distributions only assign positive probability to sets of positive Lebesgue measure, thus reflecting the idea that observations can never be taken with complete precision. Whereas this seems a reasonable assumption, it is often ignored in statistical practice, and the observations are recorded as single points, which have zero probability under the assumed sampling model. From probability theory, we know that a conditional distribution is defined up to a set of measure zero in the conditioning variable. As a consequence, the conditional distribution of the parameters given the observables (*i.e.* the posterior distribution) can fail to be well-defined for a set of measure zero in the observables.

Since any recorded sample of point observations has probability zero of occurrence, we can never be sure that the sample under consideration is not an "offending" one. What makes this problem of practical relevance is that rounding can give a nonnegligible probability to an offending sample actually being recorded. If, *e.g.* given the value zero for the observable, the posterior is not well-defined, rounding the observations to a finite precision can make it quite possible to record the value zero. Section 2 of the paper discusses these issues in some detail and presents several simple examples to illustrate this point, using both proper and improper priors.

In Section 3, we consider the situation where the conditional distribution of the parameter given the observables exists (either because a proper prior is used, or the improper prior allows for existence of such a distribution), but $p(y_0)$, the usual denominator in Bayes theorem, is infinite for the observed sample y_0. Two types of solutions are explored: to consider the limit of the posterior distribution for a sequence of values of y converging towards y_0, and to use "set observations", which identify a point observation with the neighbourhood of positive Lebesgue measure that would have led to this value being reported (*i.e.* grouped data). We show that the second solution is superior to the first one. It is, in addition, the natural approach from the point of view of probability or measure theory, since, as explained in the previous paragraph, the problems arise as a consequence of conditioning on a zero measure event (another example of the arbitrariness of conditioning on zero probability events is the famous Borel-Kolmogorov paradox). Using set observations, inference is always possible under a proper prior. For models with improper priors, we still need to verify that the mass assigned to the particular sample of set observations we consider is finite. However, once the existence of the posterior has been established for a certain sample, adding new set observations can never destroy the possibility of conducting inference. The usual coherence properties attributed to Bayesian inference are, therefore, restored through the use of set observations. Markov chain Monte Carlo methods, such as Gibbs sampling, render the solution quite feasible.

The analysis of rounded or grouped data through the use of continuous sampling models has been the object of a large literature, which was reviewed in Heitjan (1989). The focus of the latter literature has, to our knowledge, been the *quantitative* effect of coarsening on inference. This paper, on the other hand, examines the *qualitative* effect of coarsening on Bayesian inference. In other words, we deal with situations where the usual Bayesian inference is not possible on the basis of the rounded data using a continuous sampling model, and we show that explicitly incorporating the rounding mechanism is not only a natural and general, but also a feasible solution to the problem.

As a practically relevant example, Section 4 presents the case of independent sampling from scale mixtures of Normals. The Student-t distribution is an important member of this class for practical purposes. We also allow for extending these models to their skewed counterparts and complement the sampling model with a commonly used improper prior. Results are presented for the analysis using both point observations and set observations.

In Section 5, an application to stock price returns illustrates the problem and shows the empirical feasibility of the solution using set observations. This analysis is seen to be preferable to a more ad-hoc solution to the problem.

For probability density functions, we use the notation of DeGroot (1970), and all proofs are grouped in the Appendix.

2. THE FUNDAMENTAL PROBLEM

In this section we discuss the source of the problems one may face when conducting posterior inference with sampling distributions which are absolutely continuous (*i.e.* possess a density function) with respect to Lebesgue measure. For notational convenience, the prior distribution shall also be defined through a density function, but it should be noted that the problems explained here hinge in no way upon this assumption as the argument readily extends to any prior distribution.

We thus consider a sampling distribution with probability density function (p.d.f.) $p(y \mid \theta)$, with support $\mathcal{Y} \subseteq \Re^n$ and where $\theta \in \Theta \subseteq \Re^m$. We complete the Bayesian model with a σ-finite prior distribution given through a density $p(\theta)$, which could either be proper or improper. The resulting Bayesian model uniquely defines a joint σ-finite distribution on $\mathcal{Y} \times \Theta$ with density

$$p(y, \theta) = p(y \mid \theta)p(\theta). \tag{2.1}$$

2.1. *Proper Priors*

If $p(\theta)$ is proper the joint distribution defined through (2.1) can be decomposed into the marginal (predictive) distribution of y with p.d.f.

$$p(y) = \int_\Theta p(y \mid \theta)p(\theta)d\theta, \tag{2.2}$$

and the conditional distribution of θ given y, defined through the p.d.f.

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \tag{2.3}$$

if $p(y) < \infty$ and arbitrarily otherwise. Note that since $p(y)$ in (2.2) is a p.d.f., the set of y's for which $p(y) = \infty$ has Lebesgue measure zero and, thus, zero probability of being observed. However, as explained in the Introduction, current statistical practice is to conduct inference on θ on the basis of (2.3) with y replaced by the recorded point observation y_0. This can have serious implications since there is no guarantee that the reported value y_0 is not an offending value, corresponding to $p(y_0) = \infty$, in which case the usual posterior inference, based on (2.3), is precluded. The often presumed automatic feasibility of Bayesian inference under proper priors can, therefore, be destroyed by the use of point observations that are fundamentally incompatible with the sampling model.

Whenever $p(y_0)$ can be computed analytically, such a problem would be detected, but the vast majority of statistical applications to complex real-life problems has to rely on numerical methods, which may well fail to indicate the problem. Curiously, common practice does not include checking whether $p(y_0) < \infty$ in models with a proper prior. Thus, there is a danger of reporting senseless inference.

As an illustration, we present a simple example.

Example 1. *A Scale Contaminated Model*

This example concerns a scale mixture of two Normals. The behaviour of the likelihood function in a related example was studied in Kiefer and Wolfowitz (1956) and Titterington, Smith and Makov (1985, Ex. 4.3.7).

Consider n i.i.d. replications $y_1, \ldots, y_n$ from the ε-contaminated model with p.d.f.

$$p(y_i \mid \sigma) = (1 - \varepsilon) f_N(y_i \mid \mu_0, \sigma^2) + \varepsilon f_N(y_i \mid \mu_0, c^2), \tag{2.4}$$

where $\mu_0 \in \Re$ is a known quantity and $f_N(y_i \mid \xi, \omega^2)$ denotes the density function of a Normal distribution with mean ξ and variance ω^2 evaluated at y_i. In such a model $\varepsilon \in (0, 1/2)$ could represent the probability of y_i being an outlying observation, generated with variance $c^2 > 1$, whereas the usual observable has variance $\sigma^2 < 1$. For convenience, we shall assume both ε and c^2 fixed, but the following results carry over to the case with a proper prior on (ε, c^2). The prior assumed for σ will be a Beta(a, b) distribution with p.d.f.

$$p(\sigma) = B(a, b)^{-1}\sigma^{a-1}(1-\sigma)^{b-1}I_{(0,1)}(\sigma), \tag{2.5}$$

where $B(a, b)$ is the Beta function and I_H denotes the indicator function of the set H. Clearly, if $y_i \neq \mu_0$ for all $i = 1, \ldots, n$, the likelihood function from (2.4) is bounded and thus leads to a finite integral under any proper prior. If, however, $r \geq 1$ observations are equal to μ_0, the likelihood can be shown to have upper and lower bounds both proportional to σ^{-r}. Therefore, a finite predictive density value in (2.2) is achieved only when $a > r$. Thus, use of the proper prior in (2.5) with $a \leq r$ (the number of observations equal to μ_0) does not allow for posterior inference using (2.3). We also immediately see that adding new observations (equal to μ_0) can destroy the existence of a previously well-defined posterior.

The simplicity of this example, allows us to gain further insight in the source of the problem. It is immediate to see that, as $\sigma \to 0$, the sampling model in (2.4) converges in distribution to a mixture of a Dirac distribution at μ_0 (with probability $1 - \varepsilon$) and a Normal(μ_0, c^2) distribution (with probability ε). Thus, whereas any sample containing some μ_0 values has positive probability under this limiting distribution, it is a zero probability event according to the distribution in (2.4) where $\sigma > 0$. This makes the likelihood unbounded as $\sigma \to 0$, reflecting the fact that the limiting distribution is infinitely more likely to have generated such a dataset than that in (2.4). Note that a Bayesian analysis can still get around the problem if the prior for σ gives sufficiently small probability to a neighbourhood of $\sigma = 0$ [*e.g.* if a in (2.5) is larger than r].

2.2. *Improper Priors*

We now turn to the case of an improper prior $p(\theta)$. Then, the decomposition described at the beginning of Subsection 2.1 still applies if and only if the predictive distribution is σ-finite, *i.e.* the density $p(y)$ in (2.2) is finite except possibly for a set of y's of Lebesgue measure zero in $\Re^n$ [see Mouchart (1976) and Florens, Mouchart and Rolin (1990)]. Indeed, when the latter condition holds, Kolmogorov's definition of conditional distributions, which relies upon the fact that a joint distribution is absolutely continuous with respect to any of its marginals combined with the Radon-Nikodym Theorem, can directly be applied to obtain the conditional distribution of the parameter θ given the observable y. Obviously, the danger arising from plugging in a particular value y_0 in (2.3) carries over to this case. The impropriety of the prior implies that the density $p(y)$ in (2.2) can take an infinite value even if the observed sample leads to a bounded likelihood. The following example illustrates this point.

Example 2. *A Student-t Model*

Assume n independent replications from a univariate Student-t distribution with known location μ_0, unitary scale and unknown degrees of freedom ν. The likelihood function is always bounded and, as ν tends to zero, behaves like $\nu^{n-(r/2)}$, where r is the number of observations equal to μ_0. We complete the Bayesian model with an improper prior for $\nu \in (0, \infty)$, given through the density $p(\nu) \propto \nu^{-(1+a)}$, $a > 0$. From the results in the previous paragraph, we can establish the existence of a

conditional distribution of ν given the observables if and only if $n > a$, since, in this case, the marginal distribution of the observables is σ-finite. However, any observed sample for which $a \geq n - (r/2)$ leads to an infinite denominator in (2.3). Note the parallelism with Example 1, where a proper prior was used: In both cases there exists a conditional distribution of the parameters given the observables, but application of the usual formula in (2.3) is precluded for certain samples.

Finally, we present a simple model in which an improper prior performs better than a proper one.

Example 3. *Improper Prior Better Than Proper*

Suppose we have a single observation $y \in \Re$ from a Normal$(0, \sigma^2)$ distribution, and consider an Exponential prior for σ. It is then immediate that $p(y)$, computed as in (2.2), is finite if and only if $y \neq 0$. On the other hand, an improper prior with density

$$p(\sigma) = \sigma^{-1} \exp(-a\sigma^{-2}), \ a > 0, \tag{2.6}$$

implies a finite value of $p(y)$ for any $y \in \Re$. Clearly, the likelihood becomes unbounded when $y = 0$ and σ converges towards zero. The density in (2.6), which (in contrast to the Exponential) tends to zero as $\sigma \to 0$, counteracts this unboundedness. On the other hand, the lack of integrability of (2.6) as σ tends to infinity is of no consequence since the likelihood can counteract this. This simple example illustrates the fact that our usual way of reasoning (namely that a proper prior should always be safer than an improper one) does not necessarily hold if we condition on events of measure zero.

The examples presented so far are fairly straightforward, in the sense that the sampling distribution has some known fixed modal value. Thus, observations exactly equal to the mode always lead to the highest likelihood values and are the first ones that should be examined when searching for problematic samples. The much more interesting case where the mode of the sampling distribution depends on unknown parameters is more complicated to analyse. Section 4 will examine the Bayesian model corresponding to sampling from scale mixtures of Normals with unknown location and scale, under Jeffreys' prior for these parameters.

2.3. *A Remark*

Before we proceed with the remainder of the paper, we note an interesting fact. As was already mentioned in Subsection 2.1, common practice does not involve checking whether $p(y_0) < \infty$, for the observed sample y_0, in models with a proper prior. On the other hand, this is precisely the condition that is usually checked when the prior is improper. Whereas $p(y_0) < \infty$ guarantees that the expression in (2.3) with $y = y_0$ defines a p.d.f. for θ, it does not, however, imply the existence of a conditional distribution, since from $p(y_0) < \infty$ it does not follow that the predictive distribution is σ-finite. If the latter does not hold, $p(\theta \mid y_0)$ in (2.3) is properly normalized but can not be interpreted as the conditional distribution of the parameter given the observable. We can therefore mention two separate issues:

Condition A. The existence of a conditional distribution of the parameter θ given the observable y.

Condition B. The fact that (2.3) defines a p.d.f. for θ given a particular observation y_0.

Our point is that neither Condition A nor Condition B implies the other. It may well happen that A holds (under a proper prior it always does) but still $p(y_0) = \infty$ for a certain value y_0, in which case $p(\theta \mid y_0)$ in (2.3) can not be used. Conversely, the fact that $p(y_0) < \infty$ for a given value y_0 [and thus $p(\theta \mid y_0)$ in (2.3) defines a p.d.f. for θ] does not imply that a conditional distribution for θ given y exists. The ideal situation for conducting Bayesian inference is when

both A and B hold simultaneously: while A provides an interpretation of the distribution of θ given y as a conditional distribution, B seems required if we wish to conduct inference on the basis of a point observation y_0.

3. A SOLUTION THROUGH SET OBSERVATIONS

In this section we shall be concerned with situations where Condition A above holds but B does not, as we have a point observation y_0 for which $p(y_0) = \infty$. Examples 1-3 all display this behaviour.

Let us first of all examine whether taking the limit of the usual posterior distribution for a sequence of observations that converges to y_0 provides a useful solution. As a simple example, we reconsider Example 1.

Example 1. *Continued*

We now combine the sampling model in (2.4) *with the proper prior in* (2.5) *and consider* $a \leq r$*, the number of observations equal to* μ_0*. In this case* (2.2) *becomes infinite, so that the usual formula in* (2.3) *can not be applied. The limit of the posterior distributions arising from a sequence of observations that converges to the one recorded (i.e. with* r *values equal to* μ_0*), can be shown to be a Dirac distribution at* $\sigma = 0$*. This can hardly be deemed of any practical use as we would rarely be happy with concluding that the post-sample predictive (the sampling model integrated with the posterior) is outside the (continuous) class assumed in our analysis. Furthermore, such a model could provide a very inadequate fit to the data. As an example, suppose that* ε *is very small and* c^2 *very large, and that the majority of the observations are located around* μ_0 *with only one of them taking exactly the value* μ_0*. The model in* (2.4) *would then appear to be rather appropriate for some positive value of* σ*. However, if in* (2.5) *we take* $a = 1$ *and consider the limit of the posterior distributions as explained above, the post-sample predictive becomes a mixture of a Dirac distribution at* μ_0 *(with very large probability) and a Normal distribution with very large variance.*

Thus, this potential solution does not seem very satisfactory. Deriving such limiting distributions is usually quite hard and inference based on them often displays unattractive features, as illustrated in the example above. As a further example, consider independent sampling from a Normal$(0, \sigma^2)$ distribution and an Exponential prior on σ as in Example 3. If the observed values were 0 and 1, and knowledge was updated after each single observation, we would obtain different answers depending on the order in which they were observed. Indeed, observing the zero first would lead to a Dirac limiting posterior distribution for σ at zero, and any further updating would be precluded.

As a consequence of all these drawbacks, we shall not pursue this approach any further, but focus instead on solving the problem through more careful modelling of the data generating mechanism, in accordance with the way the data are actually observed.

Clearly, when a point value y_0 is recorded as an observation, we do not literally believe that y_0 is the outcome of the sampling process (indeed, it can not be according to a continuous sampling model), but it should rather be interpreted as indicative of some (small) neighbourhood S_0 around y_0. Whenever $p(y_0) = \infty$, inference will have to be based on the entire neighbourhood around y_0, rather than on the reported value alone. Thus, instead of (2.3) with $y = y_0$, we shall consider

$$p(\theta \mid y \in S_0) = \frac{P(y \in S_0 \mid \theta)p(\theta)}{P(y \in S_0)}, \tag{3.1}$$

where $P(y \in S_0 \mid \theta) = \int_{S_0} p(y \mid \theta)dy$ and $P(y \in S_0) = \int_\Theta P(y \in S_0 \mid \theta)p(\theta)d\theta$. The

crucial difference between (2.3) and (3.1) is that we now condition on an event of positive measure, namely $y \in S_0$, thus no longer contradicting the sampling assumptions. In the case of a proper prior, $p(\theta)$, this settles the issue entirely: the conditioning event has positive probability and, thus, (3.1) can immediately be used for inference on θ. If $p(\theta)$ is improper, on the other hand, we have solved the problem of conditioning on zero measure events, but we still need to check that the denominator in (3.1) is finite, so as to have a p.d.f. on θ. However, once the latter has been established for a certain sample, it can no longer be destroyed by adding new set observations.

The above procedure can be interpreted as follows: we are really observing a new random variable, say, $z = z(y)$ that takes values in a space, say $\mathcal{Z}$, of subsets of $\mathcal{Y}$ that have positive probability of occurring under $p(y \mid \theta)$. In practice, $\mathcal{Z}$ will be a countable space. In the simplest case of directly rounding the observations, the elements of $\mathcal{Z}$ will constitute a partition of $\mathcal{Y}$. A more complicated setup is where the raw data are first rounded and afterwards transformed, which implies that the sets in $\mathcal{Z}$ are not necessarily disjoint. An example of this situation will appear in Section 5. Whenever (3.1) defines a p.d.f. for θ, the counterpart of Condition B in Subsection 2.3 applies, in the sense that we can base inference on a properly normalized distribution for θ after observing $z = S_0$. Furthermore, if $\mathcal{Z}$ is countable the conditional distribution of θ given z is defined (*i.e.* the counterpart of Condition A in Subsection 2.3 holds) if and only if $P(z = S) < \infty$ for all $S \in \mathcal{Z}$. Whereas this always obtains under a proper prior, it may fail to hold if $p(\theta)$ is an improper density function.

In practice, computing (3.1) will be more complicated than (2.3), yet quite feasible through straightforward numerical methods. Note that our solution falls in the category of grouped or censored data, to which numerical methods are nowadays routinely applied, mostly through the use of data augmentation [see Tanner and Wong (1987)] on the censored observations. In particular, we can set up the simple Gibbs sampler with the following conditionals:

$$p(\theta \mid y, y \in S_0) = p(\theta \mid y), \tag{3.2}$$

$$p(y \mid \theta, y \in S_0) \propto p(y \mid \theta) I_{S_0}(y). \tag{3.3}$$

Sequential drawing from $(3.2) - (3.3)$ generates a Markov chain for $(y, \theta \mid y \in S_0)$ that will converge to the actual joint distribution and from which posterior and predictive inference can immediately be conducted. Remark that we only require the possibility to draw from the "usual" posterior p.d.f. in (2.3) and from the sampling model, truncated to the observed set S_0. In practice, an adequate pseudo-random number generator for (3.3) will never lead to offending values of y for which the predictive density is not finite, since it typically operates with high precision, so that any given value y_0 has an extremely small probability of occurrence and is very unlikely to be drawn in a run of typical length. In addition, the actual posterior would then still be well-defined, and such problems could only potentially affect the numerical issue of drawing from the conditional in (3.2).

Convergence of the Markov chain induced by $(3.2) - (3.3)$ is always guaranteed in the practically relevant case where the support of y in the sampling does not depend on θ. The latter implies that our Gibbs sampler generates a chain on $S_0 \times \Theta$ and the Cartesian product structure assures convergence as shown in Roberts and Smith (1994). For general references in the area of Markov chain Monte Carlo and Gibbs sampling, we refer the reader to Gelfand and Smith (1990), Casella and George (1992), Tierney (1994), and Besag, Green, Higdon and Mengersen (1995). Brooks (1998) provides an up-to-date review of the very extensive literature in the area of Markov chain Monte Carlo methods.

4. INDEPENDENT SAMPLING FROM SCALE MIXTURES OF NORMALS

The present section examines a leading case where Condition A in Subsection 2.3 is fulfilled for point observations, yet Condition B does not hold for certain values of the observables. In particular, we consider a location-scale model with errors that are independently distributed as scale mixtures of Normals. Thus, $y_i \in \Re$ is assumed to be generated as

$$y_i = \mu + \sigma\varepsilon_i,\ i = 1, \ldots, n, \tag{4.1}$$

with location parameter $\mu \in \Re$ and scale parameter $\sigma > 0$, where the ε_i's are i.i.d. scale mixtures of Normals with p.d.f.

$$p(\varepsilon_i \mid \nu) = \int_0^\infty f_N(\varepsilon_i \mid 0, \lambda_i^{-1})\, dP_{\lambda_i \mid \nu}, \tag{4.2}$$

for some mixing probability distribution $P_{\lambda_i \mid \nu}$ on $\Re_+$, which can depend on a parameter $\nu \in \mathcal{N}$ of finite or infinite dimension. Leading examples, which will be studied in this section are finite mixtures of Normals, with $P_{\lambda_i \mid \nu}$ a discrete distribution with finite support, Student-t sampling, with a Gamma$(\nu/2, \nu/2)$ mixing distribution, Modulated Normal type I [see Romanowski (1979)], with Pareto$(1, \nu/2)$ mixing on the support $(1, \infty)$, and Modulated Normal type II [see Rogers and Tukey (1972)], where $P_{\lambda_i \mid \nu}$ is a Beta$(\nu/2, 1)$ distribution on $(0, 1)$. A more extensive list of examples is provided in Fernandez and Steel (1996).

The parameters in the sampling model $(4.1) - (4.2)$ are (μ, σ, ν), and in the prior distribution we assume the following product structure:

$$P_{(\mu,\sigma,\nu)} = P_{(\mu,\sigma)} \times P_\nu. \tag{4.3}$$

For (μ, σ) we shall adopt the commonly used improper prior with density

$$p(\mu, \sigma) \propto \sigma^{-1}, \tag{4.4}$$

which is both the Jeffreys' prior (under "independence") and the reference prior in the sense of Berger and Bernardo (1992) when ν is known [see Fernandez and Steel (1997)]. The parameter of the mixing distribution ν will be assigned a probability measure P_ν. In a general finite mixture context with fully known components, Bernardo and Giron (1988) derive a reference prior for ν, which is then the vector of unknown mixing probabilities.

4.1. *The Analysis With Point Observations*

Here we follow common statistical practice in treating the recorded observations as values $y_1, \ldots, y_n$. For any mixing distribution $P_{\lambda_i \mid \nu}$ and any proper prior P_ν we can derive:

Result i: $p(y_1, \ldots, y_n) < \infty$ requires at least two different observations;
Result ii: if $n \geq 2$ and all observations are different, then $p(y_1, \ldots, y_n) < \infty$.

The proofs of Results i and ii proceed along the same lines as the proof of Theorem 2 (presented below). Since under a continuous sampling model the probability that any two observations are equal is zero, we can state the following result:

Theorem 1. *The Bayesian model* $(4.1) - (4.4)$ *allows for the existence of a conditional distribution of* (μ, σ, ν) *given* $(y_1, \ldots, y_n)$ *if and only if* $n \geq 2$.

Thus, Condition A of Subsection 2.3 holds for *any* scale mixture of Normals whenever we sample at least two observations. On the other hand, from Results i and ii, Condition B is not fulfilled if all the observations are equal, whereas it holds when all the observations are different. Let us now examine whether Condition B holds for samples containing some repeated observations. In this context, we obtain the following result:

Theorem 2. *Consider the Bayesian model* $(4.1)-(4.4)$ *and let* s *be the largest number of observations with the same value. If* $1 < s < n$*, we obtain* $p(y_1,\ldots,y_n) < \infty$ *if and only if*

$$\int_{0<\lambda_1\leq\ldots\leq\lambda_n<\infty} \lambda_{n-s}^{-(n-2)/2} \prod_{i\neq n-s,n} \lambda_i^{1/2}\, dP_{(\lambda_1\ldots\lambda_n)} < \infty, \tag{4.5}$$

where, with a slight abuse of notation, $P_{(\lambda_1\ldots\lambda_n)} = \int_{\mathcal{N}} (\prod_{i=1}^{n} P_{\lambda_i\,|\,\nu}) dP_\nu$.

Whereas, from Theorem 1, obtaining Condition A does not depend on the particular scale mixture of Normals considered, nor on the prior P_ν, Theorem 2 implies that both intervene when we focus on Condition B. The following theorem further examines the implications of (4.5) for some relevant examples.

Theorem 3. *Under the conditions of Theorem 2, we obtain under:*
i. *Sampling from finite mixtures of Normals:* $p(y_1,\ldots,y_n) < \infty$*;*
ii. *Student-t or Modulated Normal type II sampling:* $p(y_1,\ldots,y_n) < \infty$ *if and only if*

$$P_\nu\left(0, \frac{s-1}{n-s}\right] = 0 \; and \int_{(s-1)/(n-s)}^{\{(s-1)/(n-s)\}+\epsilon} \{(n-s)\nu - (s-1)\}^{-1} dP_\nu < \infty \; for \; all \; \epsilon > 0.$$

iii. *Modulated Normal type I sampling:* $p(y_1,\ldots,y_n) < \infty$ *if and only if*

$$P_\nu\left(0, \frac{s-1}{s}\right] = 0 \; and \int_{(s-1)/s}^{\{(s-1)/s\}+\epsilon} \{s\nu - (s-1)\}^{-1} dP_\nu < \infty \; for \; all \; \epsilon > 0.$$

Thus, Condition B is always fulfilled when sampling from finite mixtures of Normals, and the mere fact that two observations are different suffices for inference. Interestingly, inference under Student-t or Modulated Normal sampling requires bounding ν away from zero if we wish to consider samples with repeated observations. For Modulated Normal type I sampling it is sufficient to take P_ν with support on $\nu > \{(s-1)/s\} + \epsilon$ for some $\epsilon > 0$; thus, $\nu \geq 1$ always guarantees a finite predictive value. On the other hand, under Student or Modulated Normal type II models it is required that $\nu > (s-1)/(n-s)$. The latter quantity does not possess an upper bound independent of sample size, and, thus, whatever the choice of P_ν, we can always find samples for which $p(y_1,\ldots,y_n) = \infty$. In practice, one often chooses a prior for ν with support on all of $\Re_+$, which means that the problem will appear under Student-t or Modulated Normal sampling as soon as two observations in the sample are equal. A very disturbing consequence is that adding new observations can actually destroy the existence of a posterior that was perfectly well-defined with the previous sample!

4.2. *The Analysis With Set Observations*

Let us now apply the solution proposed in Section 3 to the model $(4.1)-(4.4)$. Thus, instead of point observations, we shall consider as our data information that $y_i \in S_i, i = 1,\ldots,n$, where S_i is a neighbourhood of y_i. Since the prior assumed in $(4.3)-(4.4)$ is not proper, we need to verify whether $P(y_1 \in S_1,\ldots,y_n \in S_n) < \infty$ before inference can be conducted. The following theorem addresses this issue.

Theorem 4. *Consider the Bayesian model* $(4.1)-(4.4)$ *with any mixing distribution* $P_{\lambda_i\,|\,\nu}$ *and any proper prior* P_ν*. The observations consist of* n *bounded intervals* $S_1,\ldots,S_n$ *(of*

positive Lebesgue measure in $\Re$*). Then* $P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$ *if and only if* $n \geq 2$ *and there exist two sets, say,* S_i *and* S_j *for which*

$$\inf_{y_i \in S_i, y_j \in S_j} \mid y_i - y_j \mid > 0. \tag{4.6}$$

Thus, the existence of at least two intervals that are strictly separated from each other is a necessary and sufficient condition for inference on the basis of these set observations. The necessity of this condition can be seen as the set counterpart of Result i in Subsection 4.1. Now, however, this condition is also sufficient for inference with any scale mixture of Normals. Thus, irrespective of the mixing distribution and the prior P_ν, the counterpart of Condition B for set observations always holds under (4.6), whereas we know that it fails for any sample not satisfying (4.6). On the other hand, the counterpart of Condition A will now never obtain since the collection of offending values, *i.e.* the samples of sets not verifying (4.6), has positive probability of being observed. Nevertheless, this does not preclude inference on the basis of any sample of set observations for which (4.6) holds, as is most likely in practice.

4.3. *Skewed Scale Mixtures of Normals*

In some situations the symmetry assumption implicit in the model $(4.1) - (4.2)$ might be considered inappropriate for the data at hand. In such cases, we can follow the proposal of Fernandez and Steel (1998) in order to introduce skewness into the model. In particular, we can replace the density of the error term in (4.2) by

$$p(\varepsilon_i \mid \nu, \gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \int_0^\infty \left\{ f_N\left(\frac{\varepsilon_i}{\gamma} \mid 0, \lambda_i^{-1}\right) I_{[0,\infty)}(\varepsilon_i) + f_N\left(\gamma\varepsilon_i \mid 0, \lambda_i^{-1}\right) I_{(-\infty,0)}(\varepsilon_i) \right\} dP_{\lambda_i \mid \nu}, \tag{4.7}$$

where ν is as before and we introduce a parameter $\gamma \in \Re_+$. Thus, (4.7) is obtained from (4.2) by scaling with γ to the right of the origin and with its inverse to the left of zero. Clearly, for $\gamma = 1$ (4.7) coincides with (4.2), but if $\gamma > 1$ we introduce right skewness, whereas values of $\gamma < 1$ lead to left skewed distributions. More details on the properties of such distributions are provided in Fernandez and Steel (1998).

The prior distribution is now given by

$$P_{(\mu,\sigma,\nu,\gamma)} = P_{(\mu,\sigma)} \times P_\nu \times P_\gamma, \tag{4.8}$$

where $P_{(\mu,\sigma)}$ is described in (4.4) and P_ν and P_γ are any probability measures.

The following result addresses the influence of our skewness transformation.

Theorem 5. *Consider the Bayesian model* (4.1), (4.4), $(4.7) - (4.8)$.

i. With point observations $y_1, \ldots, y_n$ *we obtain* $p(y_1, \ldots, y_n) < \infty$ *if and only if the same holds when* $\gamma = 1$.

ii. With set observations $S_1, \ldots, S_n$*, we obtain* $P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$ *if and only if the same holds when* $\gamma = 1$.

Surprisingly, the extra flexibility in dealing with skewness does not affect the possibility of conducting inference, although the actual numerical results might, of course, be quite different. Thus, all results presented in Subsections 4.1 and 4.2 for the symmetric model immediately apply to the skewed case. The next section will present an application of skewed Student sampling to a financial data set.

5. AN APPLICATION TO STOCK PRICE RETURNS

The data we will examine here were taken from Buckle (1995), and represent a sample of 49 returns on Abbey National shares between July 31 and October 8, 1991. These returns are constructed from price data p_i, $i = 0, \ldots, 49$, as $y_i = (p_i - p_{i-1})/p_{i-1}$, $i = 1, \ldots, 49$. As the data seem to exhibit some skewness, Buckle (1995) uses a Stable distribution, allowing for asymmetry. Here, we shall follow Fernandez and Steel (1998) and use instead the skewed Student sampling model obtained from (4.1) and (4.7) with $P_{\lambda_i \mid \nu}$ a Gamma$(\nu/2, \nu/2)$ distribution, which we combine with the prior distribution in (4.8) where $P_{(\mu,\sigma)}$ is as described in (4.4). In this particular application, we shall choose an exponential prior distribution for ν with mean 10 and variance 100, spreading the prior mass over a wide range of tail behaviour, and a Normal$(0, \pi/2)$ distribution truncated to $\Re_+$ for γ. The latter centers the prior over $\gamma = 1$, *i.e.* symmetry, and provides a compromise between sufficient spread and reasonably equal prior weights to right and left skewness.

Let us first consider the analysis with point observations: Theorem 1 assures us that Condition A in Subsection 2.3 holds as $n \geq 2$. However, the data contain seven observations that are recorded as zero. Thus, from Theorem 3 (ii) we know that Condition B does not hold with this data set, since $(s-1)/(n-s) = 6/42 = 1/7$ and the prior distribution for ν has mass arbitrarily close to zero. Bayesian inference on the basis of this sample is, therefore, precluded. This problem was avoided in Fernandez and Steel (1998) by slightly perturbing the original data, thus avoiding repeated observations. However, this solution is arbitrary and not in accordance with the way the data are recorded. Here we will, instead, consider the solution proposed in Section 3.

The set observations corresponding to this sample are constructed as follows: prices were recorded in integer values (in Pence) and we shall assume they were rounded to the nearest integer. The set observations for the returns are then defined as

$$S_i = \left(\frac{p_i - p_{i-1} - 1}{p_{i-1} + 0.5}, \frac{p_i - p_{i-1} + 1}{p_{i-1} - 0.5} \right), \tag{5.1}$$

$i = 1, \ldots, 49$. As a consequence of the return transformation after rounding the prices, the sets S_i are not all pairwise disjoint, yet we can find at least two sets for which (4.6) holds. Thus, Bayesian inference on the basis of set observations is possible from Theorems 4 and 5.

The numerical analysis will be conducted as indicated in Section 3. In this particular model, data augmentation with the mixing parameters $\lambda_1, \ldots, \lambda_n$ will facilitate the Gibbs sampler used for the posterior analysis. Thus, the complete Gibbs sampler will be conducted on $(y_1, \ldots, y_n, \mu, \sigma, \nu, \gamma, \lambda_1, \ldots, \lambda_n)$. For the full conditionals of μ, σ, ν, γ and $(\lambda_1, \ldots, \lambda_n)$ we refer the reader to Fernandez and Steel (1998). Whereas the latter constitute (3.2), we now need to add the full conditional distribution of $(y_1, \ldots, y_n)$ [*i.e.* (3.3)], which is a product of n skewed Normal distributions truncated to the set observations. In all, the Gibbs sampler generates a Markov chain in $2n + 4$ dimensions by cycling through six steps.

The continuous lines in Figures 1-4 display the posterior p.d.f.'s of $\mu, \tau = \sigma^{-1}, \gamma$ and ν for the set observations in (5.1). Some evidence for right skewness transpires from Figure 3 as values for $\gamma > 1$ receive most of the posterior mass. The data also indicate some support for relatively thick tails (Figure 4), although the small data set under consideration is not very informative on tail behaviour.

Let us now contrast this analysis with the one based on perturbed point observations. The perturbation was applied to the price data, p_i, and consisted in adding a Uniformly distributed random number on $(-5 \times 10^{-7}, 5 \times 10^{-7})$ to the recorded prices, who are themselves of the order 300. For a given perturbation, the resulting point observations $y_i = (p_i - p_{i-1})/p_{i-1}$, $i =$

$1, \ldots, 49$ no longer contained any repeated values and dashed lines in Figures 1-4 summarize posterior inference. As indicated by the present empirical evidence, the choice between set observations and a small ad-hoc perturbation need not be a major issue in cases where the problematic area receives very little posterior mass. We remind the reader that problems occur for the original unperturbed point observations whenever $\nu \leq 1/7$. As Figure 4 shows, very little posterior probability is allocated to that region for ν. Since the Markov chain is unlikely to wander in this area, the particular solution adopted need not make a large difference in this case. If we force the issue, however, and fix ν at a problematic value, say $\nu = 0.1 < 1/7$, we observe a very different picture.

Clearly, the tails of the Student-t sampling model with $\nu = 0.1$ are too thick to adequately fit this data set, which displays quite a concentration of mass around the mode. As a consequence, the model will try to accommodate the empirical mass around the mode by increasing the precision $\tau = \sigma^{-1}$. Thus, the observations that are not close to the mode will tend to be regarded as "outliers" with relatively small weights (*i.e.* small values of the mixing variable λ_i) attached to them. This happens both when set observations are used and with perturbed data. However, the degree to which this phenomenon affects the results is quite different.

Figures 5-7 graphically display the posterior p.d.f.'s of μ, $\ln(\tau)$ and γ. Let us first comment on the results using set observations; as expected, the precision, τ, has its mass at much higher values than in the case with free ν [note we now graph $\ln(\tau)$]. As the precision is so high, the posterior of μ will switch between the local modes in the empirical distribution of the complete data $y_1, \ldots, y_n$, depending on how they are drawn in their intervals $S_1, \ldots, S_n$. This strange behaviour of μ should be a clear warning to the practitioner that the model with $\nu = 0.1$ is not a good choice for this data set. The inference on the skewness parameter, γ, is surprisingly little affected by the restriction on ν.

If we use perturbed point observations, the concentration of the data around zero is much higher: whereas the seven repeated observations roughly lie in the set $(-0.033, 0.033)$ if we use set observations, the corresponding perturbed point observations are all situated in the interval $(-2 \times 10^{-9}, 2 \times 10^{-9})$. This translates into a much higher precision, evident from Figure 6. Virtually all the weight is now assigned to the seven perturbed zero observations (λ_i's of the order 10), whereas the 42 remaining observations are practically discarded (λ_i's of the order 10^{-11}). As a consequence, μ gets almost all of its mass very close to zero (Figure 5). In addition, Figure 7 no longer displays evidence of right skewness in the data (it so happens that the perturbed observations are somewhat bunched on the negative axis, and five out of the seven are situated to the left of the posterior mean on μ).

Clearly, when we move to more dangerous waters by imposing ν equal to a value for which the original point observations do not allow for inference, the issue of how this problem is resolved becomes of critical importance. We run into problems if we use small ad-hoc perturbations, whereas larger perturbations risk seriously biasing the inference. The only real solution to the problem seems, in our view, to be through a coherent use of set observations.

ACKNOWLEDGEMENTS

We gratefully acknowledge stimulating discussions with Peter Green, John Hartigan, Michael Lavine, Dale Poirier, Jean-Francois Richard, Richard L. Smith and Harald Uhlig. Part of this research was carried out at the Statistics Department of Purdue University, facilitated by a travel grant of the Netherlands Organization for Scientific Research (NWO). Both authors were affiliated to CentER, Tilburg University, during much of the work on this paper. The first author was supported by a Training and Mobility of Researchers fellowship (ERBFMBICT 961021), financed by the European Commission.

REFERENCES

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

Bernardo, J.M. and Giron, F.J. (1988). A Bayesian analysis of simple mixture problems (with discussion). *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 67–78.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10**, 3–66.

Brooks, S.P. (1998). Markov chain Monte Carlo and its application. *The Statistician* **47**, 1–33.

Buckle, D.J. (1995). Bayesian inference for stable distributions. *J. Amer. Statist. Assoc.* **90**, 605–613.

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46**, 167–174.

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

Fernandez, C., and Steel, M.F.J. (1996). On Bayesian inference under sampling from scale mixtures of Normals. *Tech. Rep.* **9602**, CentER, Tilburg University, The Netherlands.

Fernandez, C., and Steel, M.F.J. (1997). Reference priors for the general location-scale model. *Tech. Rep.* **97105**, CentER, Tilburg University, The Netherlands.

Fernandez, C., and Steel, M.F.J. (1998). On Bayesian modelling of fat tails and skewness. *J. Amer. Statist. Assoc.* **93**, 359–371.

Florens, J.P., Mouchart, M. and Rolin, J.M. (1990). Invariance arguments in Bayesian statistics. *Economic Decision Making: Games, Econometrics and Optimisation*. (J. Gabszewicz, J.F. Richard and L.A. Wolsey, eds.). Amsterdam: North-Holland, 387–403.

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

Heitjan, D.F. (1989). Inference from grouped continuous data: A review. *Statist. Sci.* **4**, 164–183.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887–906.

Mouchart, M. (1976). A note on Bayes theorem. *Statistica* **36**, 349–357.

Roberts, G.O., and Smith, A.F.M. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stoch. Processes Applicat.* **49**, 207–216.

Rogers, W.H. and Tukey, J.W. (1972). Understanding some long-tailed symmetric distributions. *Statistica Neerlandica* **26**, 211–226.

Romanowski, M. (1979). *Random Errors in Observation and the Influence of Modulation on Their Distribution*. Stuttgart: Konrad Wittwer.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82**, 528–550.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701–1762.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

APPENDIX: PROOFS OF THE THEOREMS

Proof of Theorem 2

Consider the joint distribution of $(y_1, \ldots, y_n, \lambda_1, \ldots, \lambda_n, \mu, \sigma, \nu)$ corresponding to $(4.1) - (4.4)$. After integrating out μ with a Normal distribution, σ through a Gamma distribution on σ^{-2}, and ν, we are left with

$$p(y_1, \ldots, y_n) \propto \int_{\Re_+^n} \left(\prod_{i=1}^{n} \lambda_i^{1/2} \right) \left(\sum_{i=1}^{n} \lambda_i \right)^{(n-2)/2} S^2(\lambda, y)^{-(n-1)/2} dP_{(\lambda_1, \ldots, \lambda_n)}, \qquad (A.1)$$

where

$$S^2(\lambda, y) = \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j (y_i - y_j)^2. \qquad (A.2)$$

Note that $\sum_{i=1}^n \lambda_i$ has upper and lower bounds which are both proportional to the biggest λ_i, whereas $S^2(\lambda, y)$ has upper and lower bounds proportional to the biggest product $\lambda_i \lambda_j$ for which $y_i \neq y_j$. Thus, Theorem 2 follows.

Proof of Theorem 3

From Theorem 2, we need to check whether (4.5) is fulfilled for each of the sampling distributions considered. This is immediate for finite mixtures of Normals. We prove parts ii and iii by making use of Fubini's theorem, computing (4.5) as

$$\int_0^\infty I(\nu) dP_\nu, \qquad (A.3)$$

with

$$I(\nu) = \int_{0 < \lambda_1 \leq \ldots \leq \lambda_n < \infty} \lambda_{n-s}^{-(n-2)/2} \prod_{i \neq n-s,n} \lambda_i^{1/2} dP_{\lambda_1 \,|\, \nu} \ldots dP_{\lambda_n \,|\, \nu}, \qquad (A.4)$$

where $P_{\lambda_i \,|\, \nu}$ is the mixing distribution corresponding to each of the sampling models.

Under Student-t sampling, $P_{\lambda_i \,|\, \nu}$ is a Gamma$(\nu/2, \nu/2)$ distribution. For $\lambda_1 \leq \ldots \leq \lambda_n$, $\prod_{i \neq n-s,n} \lambda_i^{1/2} \leq \lambda_{n-s}^{(n-s-1)/2} \lambda_n^{(s-1)/2}$, and we can prove that $I(\nu)$ is bounded (and, thus, integrable with respect to P_ν) as $\nu \to \infty$. Let us now consider a bounded interval, say $(0, B)$, for ν. Making use of the bounds

$$\frac{w^v}{v} \exp(-rw) \leq \int_0^w \lambda^{v-1} \exp(-r\lambda) d\lambda \leq \frac{w^v}{v}, \text{ for any } r, v, w > 0, \qquad (A.5)$$

while iteratively integrating $\lambda_1, \ldots, \lambda_n$, shows that $I(\nu) < \infty$ requires $\nu > (s-1)/(n-s)$. In this case, $I(\nu)$ has an upper and a lower bound which are both proportional to $\{(n-s)\nu - (s-1)\}^{-1}$. This immediately leads to Theorem 3 (ii) for Student-t sampling. The proofs for Modulated Normal type I and type II sampling are simplified versions of this one, since $I(\nu)$ in $(A.4)$ can be computed directly.

Proof of Theorem 4

After integrating out μ, σ and ν from the joint distribution of $(y_1, \ldots, y_n, \lambda_1, \ldots, \lambda_n, \mu, \sigma, \nu)$, which requires $n \geq 2$, we are left with $p(y_1, \ldots, y_n)$ in $(A.1)$, which still needs to be integrated over the sets $S_1, \ldots, S_n$. Applying Fubini's theorem, we shall first perform the integral over these sets, dealing with the integral with respect to $P_{(\lambda_1, \ldots, \lambda_n)}$ afterwards. Thus, we are first concerned with evaluating

$$T(\lambda) = \int_{S_1 \times \ldots \times S_n} S^2(\lambda, y)^{-(n-1)/2} dy_1 \ldots dy_n, \qquad (A.6)$$

with $S^2(\lambda, y)$ defined in $(A.2)$.

<u>*Sufficiency:*</u> Let us assume that (4.6) holds for, say, S_1 and S_2. First observe that

$$S^2(\lambda, y) = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} (\sum_{i=1}^n \lambda_i) \eta_2^2 + (\eta_3 - \rho, \ldots, \eta_n - \rho) Q (\eta_3 - \rho, \ldots, \eta_n - \rho)', \qquad (A.7)$$

where $\eta_i = y_1 - y_i$ for $i = 2, \ldots, n$, $\rho = \lambda_2\eta_2/(\lambda_1 + \lambda_2)$ and $Q = (q_{ij})_{i,j=3}^n$, with diagonal elements $q_{ii} = \lambda_i \sum_{j\neq i} \lambda_j$, and off-diagonal elements $q_{ij} = q_{ji} = -\lambda_i\lambda_j$. Since, by assumption, $\mid \eta_2 \mid \geq K$ for some constant $K > 0$, the integrand in $(A.6)$ is the kernel of a Cauchy distribution for $(\eta_3, \ldots, \eta_n)'$. Making a transformation from $y_1, \ldots, y_n$ to $y_1, \eta_2, \ldots, \eta_n$ and integrating $(\eta_3, \ldots, \eta_n)'$ over the whole of $\Re^{n-2}$ using the latter Cauchy distribution, leads to

$$T(\lambda) \leq (\prod_{i=1}^n \lambda_i^{-1/2})(\sum_{i=1}^n \lambda_i)^{-(n-2)/2} \int_{\{y_1\in S_1, y_1-\eta_2\in S_2\}} \mid \eta_2 \mid^{-1} dy_1 d\eta_2. \qquad (A.8)$$

This integral is finite since S_1 and S_2 are bounded and $\mid \eta_2 \mid \geq K > 0$. Combining $(A.1)$, $(A.6)$ and $(A.8)$ immediately implies that $P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$ under any probability measure $P_{(\lambda_1,\ldots,\lambda_n)}$.

Necessity: Defining $\eta_2, \ldots, \eta_n$ as before, we have $S^2(\lambda, y) = \eta'\tilde{Q}\eta$, where $\eta = (\eta_2, \ldots, \eta_n)'$ and $\tilde{Q} = (q_{ij})_{i,j=2}^n$ with the elements q_{ij} defined in the same way as the elements of Q in $(A.7)$. Since $\tilde{Q}$ is a PDS matrix, it can be expressed as $\tilde{Q} = D'D$ for some $(n-1)\times(n-1)$ nonsingular matrix D. We consider a variable transformation from $y_1, \ldots, y_n$ to $y_1, \xi_2, \ldots, \xi_n$, where $\xi = (\xi_2, \ldots, \xi_n)' = D\eta$. If (4.6) does not hold, the image set of $S_1 \times \ldots \times S_n$ in the transformed variables will contain an $(n-1)$-dimensional connected set, C, for ξ, the closure of which contains the $(n-1)$-dimensional vector of zeroes. This leads to

$$T(\lambda) \geq \mid D \mid^{-1} \int_{S_1} dy_1 \int_C (\sum_{i=2}^n \xi_i^2)^{-(n-1)/2} d\xi_2 \ldots d\xi_n.$$

The last integral is seen to be infinite after a polar transformation.

Proof of Theorem 5

From the unimodality of the Normal distribution, the following upper and lower bounds for $p(\varepsilon_i \mid \nu, \gamma)$ in (4.7) can be derived

$$\frac{2}{\gamma + \frac{1}{\gamma}} \int_0^\infty f_N\left(\frac{\varepsilon_i}{h(\gamma)} \mid 0, \lambda_i^{-1}\right) dP_{\lambda_i \mid \nu}, \quad h(\gamma) = \begin{cases} \max\{\gamma, 1/\gamma\} & \text{for upper bound} \\ \min\{\gamma, 1/\gamma\} & \text{for lower bound.} \end{cases} \qquad (A.9)$$

This allows us to bound $p(y_1, \ldots, y_n)$ by

$$2^n \int (\prod_{i=1}^n \lambda_i^{1/2}) \frac{\sigma^{-(n+1)}}{(\gamma + \frac{1}{\gamma})^n} \exp\left\{-\frac{1}{2\sigma^2 h(\gamma)^2} \sum_{i=1}^n \lambda_i (y_i - \mu)^2\right\} d\mu d\sigma dP_\gamma dP_{(\lambda_1,\ldots,\lambda_n)}, \quad (A.10)$$

which, after transforming from σ to $\vartheta = h(\gamma)\sigma$, can be rewritten as

$$2^n \int \left(\frac{h(\gamma)}{\gamma + \frac{1}{\gamma}}\right)^n dP_\gamma \int (\prod_{i=1}^n \lambda_i^{1/2}) \vartheta^{-(n+1)} \exp\left\{-\frac{1}{2\vartheta^2} \sum_{i=1}^n \lambda_i (y_i - \mu)^2\right\} d\mu d\vartheta dP_{(\lambda_1,\ldots,\lambda_n)}. \qquad (A.11)$$

Clearly, for both choices of $h(\gamma)$, the value of the first integral in $(A.11)$ lies in the interval $(0, 1)$ under any proper prior P_γ. On the other hand, the second integral in $(A.11)$ corresponds to $p(y_1, \ldots, y_n)$ when γ is fixed at the value 1. This proves Theorem 5.

Figure 1: Posterior Density for mu

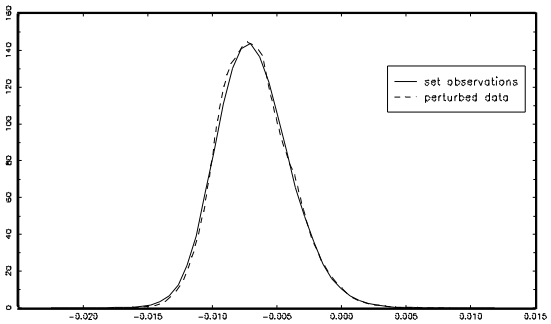


Figure 2: Posterior Density for tau

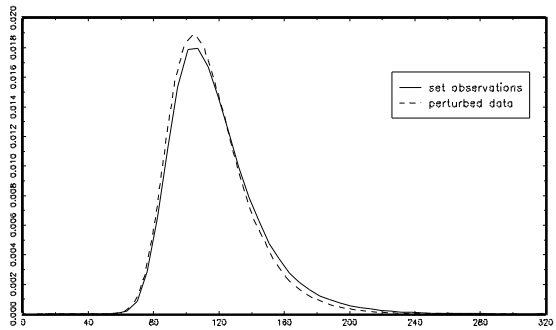


Figure 3: Posterior Density for gamma

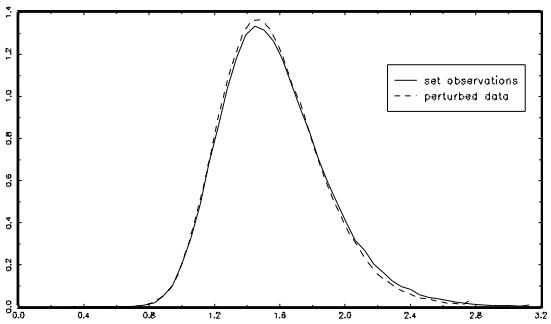


Figure 4: Posterior Density for nu

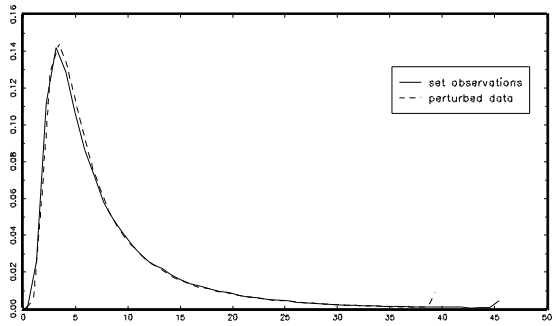


Figure 5: Posterior Density for mu, nu=0.1

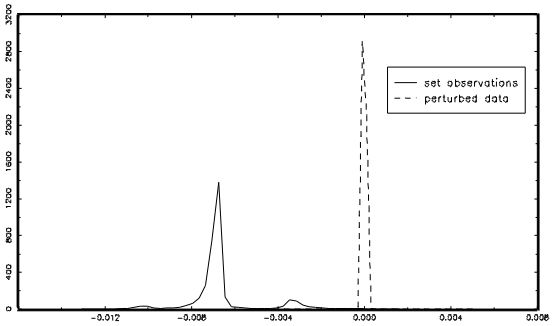


Figure 6: Posterior Density for ln(tau), nu=0.1

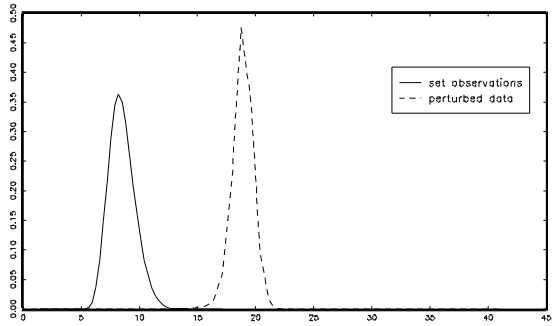
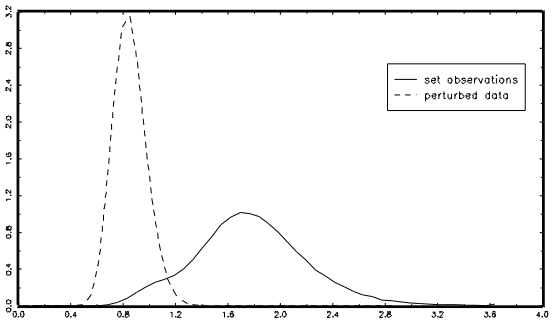


Figure 7: Posterior Density for gamma, nu=0.1



Figures 1-7. *Posterior results for stock price returns*