Statistics and Computing (1991) **1**, 105–117

Bayesian analysis of outlier problems using the Gibbs sampler

ISABELLA VERDINELLI[1,2] and LARRY WASSERMAN[2]

[1]*Department of Statistics, University of Rome, Piazzale A. Moro 5,00185 Rome, Italy*

[2]*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA*

Received September 1990 and accepted November 1990

We consider the Bayesian analysis of outlier models. We show that the Gibbs sampler brings considerable conceptual and computational simplicity to the problem of calculating posterior marginals. Although other techniques for finding posterior marginals are available, the Gibbs sampling approach is notable for its ease of implementation. Allowing the probability of an outlier to be unknown introduces an extra parameter into the model but this turns out to involve only minor modification to the algorithm. We illustrate these ideas using a contaminated Gaussian distribution, a t-distribution, a contaminated binomial model and logistic regression.

Keywords: Contaminated normal, Gibbs sampling, Monte Carlo, outliers, posterior marginals

1. Introduction

Calculating posterior marginals for an outlier model typically involves difficult computations. The simplest example of this type of model is a finite mixture of normal distributions. Box and Tiao (1968), Guttman *et al* (1978), Abraham and Box (1978), Freeman (1980), Pettit and Smith (1984; 1985) and Titterington *et al* (1985) consider such models. All these authors comment on the computational problems in dealing with these models. For example, Freeman (1980), commenting on the work of Box and Tiao, Abraham and Box, Guttman *et al*, says that 'all three models can only be used with small sample sizes unless a maximum of two outliers is contemplated'.

We will show how for such models, Bayesian statistical analysis is simple using a Monte Carlo technique known as the Gibbs sampler (Geman and Geman, 1984; Tanner and Wong, 1987; Gelfand and Smith, 1990). Furthermore, we shall be able to carry out a fully Bayesian analysis. That is, we shall not assume that the number of outliers, or the probability that an observation is an outlier, is known. These simplifications are not needed when using the Gibbs algorithm.

An outlier is usually defined to be an observation that does not come from the assumed model or an extreme observation that is far away from the rest of the observations. Giving a precise definition to the concept of an outlier is difficult since the notion of an 'extreme observation' is subtle. We shall not attempt a rigorous definition here. We refer the reader to Pettit and Smith (1985) for a discussion. An alternative view is presented in Chaloner and Brant (1988).

The simplest, and perhaps most studied, case is the normal model with mean μ and variance σ^2. To allow for the possibility of outliers, the model is enhanced so that the density for the observation y_i is of the form

$$f(y_i \mid \mu, \sigma^2, \epsilon, A_i) = (1-\epsilon)\phi(y_i \mid \mu, \sigma^2) + \epsilon\phi(y_i \mid \mu + A_i, \sigma^2)$$

Here, $\phi(y \mid \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2, and $\epsilon \in [0, 1]$ is the probability that the ith observation is from a normal model whose location is shifted by a factor A_i. This is known as the *contaminated* (location-shift) normal. Even if ϵ is assumed known, this model is cumbersome since the analysis depends on the number of outliers in the model. Guttman *et al* (1978) assume that the number of outliers k is known and, further, that each subset of k observations is equally likely to be a set of outliers. Even so, the computations are still difficult.

Usually, the posterior marginal for μ is the main concern. Typically, we would also like to compute, for each observation, the posterior probability of that observation being an outlier. To be realistic, we should treat ϵ as unknown as well. Doing so increases the dimension of the

problem. Furthermore, the likelihood is difficult to work with since it is the product of mixtures. This makes finding posterior probabilities considerably more difficult. We shall see that the problem is well suited to the Gibbs sampler. Our main point is not that other techniques will not work, but that the Gibbs sampler brings striking conceptual simplicity to the computations. It is simplicity, not efficiency, that is the biggest obstacle to the implementation of Bayesian methods and this paper attempts to show how successful the Gibbs sampler is for obtaining such simplicity.

The reason why the Gibbs sampler brings such simplicity to outlier problems is that the Gibbs sampler operates by iterating two different stages of calculations. In the first stage, we assume we know which observations are outliers. This allows us to correct the outliers, and then we sample from the posterior using the corrected data. In the second stage, we assume that we know the true parameter values and, for each observation, we compute the posterior probability that the observation is an outlier. This is simple since the parameters are assumed to be known. Alternating the two stages will eventually allow us to draw a sample from the posterior distribution. The details will be made clear in what follows. The important point is that the Gibbs sampler divides a difficult problem into a set of simpler problems.

The outline of this paper is as follows. Section 2 gives a brief description of the Gibbs sampler. In Section 3 we treat the contaminated normal location-scale problem. In Section 4, we consider using the student t-distribution as a model for the sampling distribution. This distribution has been suggested as an alternative to the contaminated normal as a way of modeling the fact that extreme observations are possible (Fraser, 1979; West, 1987; Lange *et al*, 1989).

Dealing with outliers in binomial problems (Winkler and Gaba, 1990) is also manageable, as we show in Section 5. We extend this to binomial regression (Pregibon, 1981; 1982; Copas, 1988) as well. Copas notes that, at most, a profile likelihood for each observation being an outlier is the best that can be obtained from the likelihood analysis. Here, we obtain the posterior probability that each observation is an outlier.

Section 6 contains a discussion of the results and indicates how the methods described in this paper can easily be generalized to handle regression problems and to handle outliers in other models such as exponential distributions (Pettit, 1988).

2. Gibbs sampling

In this section we give a short description of the Gibbs sampling algorithm, as considered in Gelfand and Smith (1990). What follows is the most basic form of the algorithm, and we do not examine all possible variations. More details can be found in Gelfand and Smith (1990). The algorithm is intimately related to the notion of data augmentation and substitution sampling (Tanner and Wong, 1987).

We consider, for illustration, the case of three parameters only $(\theta_1, \theta_2, \theta_3)$, and we assume that the three full conditional posterior distributions $f_1(\theta_1 | \theta_2, \theta_3, \mathbf{y})$, $f_2(\theta_2 | \theta_1, \theta_3, \mathbf{y})$ and $f_3(\theta_3 | \theta_1, \theta_2, \mathbf{y})$ are available, meaning only that random samples can be drawn from them. Here, $\mathbf{y}$ denotes a vector $(y_1, y_2, \ldots, y_n)$ of n observations. If f_i is known in closed form and is a familiar distribution, then standard routines are available for drawing random numbers from f_i. If f_1, say, is not available in closed form, then one can obtain f_1 up to a proportionality constant by evaluating the product of the likelihood and the prior over a grid of values for θ_1, with θ_2, θ_3 and $\mathbf{y}$ fixed. Then, standard numerical techniques can be used to generate an observation from f_1 without renormalizing the product of the likelihood and prior. This is straightforward since θ_1 is one-dimensional. The simplest technique is probably the rejection sampling method (DeVroye, 1986). In its crudest form, one samples x uniformly on the support of f_1 (assuming the support is compact) and then samples w uniformly from 0 to m, where m is the maximum of $k(\theta_1)$. Here, $k(\theta_1)$ is the un-normalized product of the likelihood and prior (with θ_2, θ_3 and $\mathbf{y}$ fixed) so that $f_1 = k/\int k$. If $w \leq k(x)$ we keep x, otherwise we throw away x and draw a new x and a new w. We continue until we keep an x. The value that results is a random draw from f_1. This can be a very inefficient way to generate random draws from f_1 and many refinements are possible to make the process more efficient.

The Gibbs sampler is a simple way to generate observations from the joint posterior distribution so that the posterior marginal densities $f(\theta_1 | \mathbf{y})$, $f(\theta_2 | \mathbf{y})$ and $f(\theta_3 | \mathbf{y})$ can be estimated. To describe the algorithm, we begin by considering R groups of arbitrary starting values for the three parameters θ_i, $i = 1, 2, 3$:

$$\{(\theta_1)_0^1 (\theta_2)_0^1 (\theta_3)_0^1\}, \{(\theta_1)_0^2 (\theta_2)_0^2 (\theta_3)_0^2\}, \ldots,$$
$$\{(\theta_1)_0^r (\theta_2)_0^r (\theta_3)_0^r\}, \ldots, \{(\theta_1)_0^R (\theta_2)_0^R (\theta_3)_0^R\}$$

From each of these R groups of starting values we generate S sets of random numbers drawn from the conditional posterior distributions above. More specifically, consider the rth group—the first set of random numbers, $\{(\theta_1)_1^r, (\theta_2)_1^r, (\theta_3)_1^r\}$, is obtained as follows:

$$(\theta_1)_1^r \text{ is drawn from } f_1(\theta_1 | (\theta_2)_0^r, (\theta_3)_0^r, \mathbf{y})$$
$$(\theta_2)_1^r \text{ is drawn from } f_2(\theta_2 | (\theta_1)_1^r, (\theta_3)_0^r, \mathbf{y})$$
$$(\theta_3)_1^r \text{ is drawn from } f_3(\theta_3 | (\theta_1)_1^r, (\theta_2)_1^r, \mathbf{y})$$

The second set of random numbers, $\{(\theta_1)_2^r, (\theta_2)_2^r, (\theta_3)_2^r\}$, is

obtained as follows:

$$(\theta_1)_2^r \text{ is drawn from } f_1(\theta_1 \mid (\theta_2)_1^r, (\theta_3)_1^r, \mathbf{y})$$
$$(\theta_2)_2^r \text{ is drawn from } f_2(\theta_2 \mid (\theta_1)_2^r, (\theta_3)_1^r, \mathbf{y})$$
$$(\theta_3)_2^r \text{ is drawn from } f_3(\theta_3 \mid (\theta_1)_2^r, (\theta_2)_2^r, \mathbf{y})$$

The procedure is repeated S times to generate the following collection of random numbers:

$$\begin{vmatrix} (\theta_1)_1^r & (\theta_2)_1^r & (\theta_3)_1^r \\ (\theta_1)_2^r & (\theta_2)_2^r & (\theta_3)_2^r \\ \vdots & \vdots & \vdots \\ (\theta_1)_s^r & (\theta_2)_s^r & (\theta_3)_s^r \\ \vdots & \vdots & \vdots \\ (\theta_1)_S^r & (\theta_2)_S^r & (\theta_3)_S^r \end{vmatrix}$$

The above step is repeated for $r = 1, \ldots, R$ to obtain the following R collections of S sets of random numbers:

$$\begin{vmatrix} (\theta_1)_1^1 & (\theta_2)_1^1 & (\theta_3)_1^1 \\ (\theta_1)_2^1 & (\theta_2)_2^1 & (\theta_3)_2^1 \\ \vdots & \vdots & \vdots \\ (\theta_1)_s^1 & (\theta_2)_s^1 & (\theta_3)_s^1 \\ \vdots & \vdots & \vdots \\ (\theta_1)_S^1 & (\theta_2)_S^1 & (\theta_3)_S^1 \end{vmatrix}, \quad \begin{vmatrix} (\theta_1)_1^2 & (\theta_2)_1^2 & (\theta_3)_1^2 \\ (\theta_1)_2^2 & (\theta_2)_2^2 & (\theta_3)_2^2 \\ \vdots & \vdots & \vdots \\ (\theta_1)_s^2 & (\theta_2)_s^2 & (\theta_3)_s^2 \\ \vdots & \vdots & \vdots \\ (\theta_1)_S^2 & (\theta_2)_S^2 & (\theta_3)_S^2 \end{vmatrix}, \ldots$$

$$\begin{vmatrix} (\theta_1)_1^r & (\theta_2)_1^r & (\theta_3)_1^r \\ (\theta_1)_2^r & (\theta_2)_2^r & (\theta_3)_2^r \\ \vdots & \vdots & \vdots \\ (\theta_1)_s^r & (\theta_2)_s^r & (\theta_3)_s^r \\ \vdots & \vdots & \vdots \\ (\theta_1)_S^r & (\theta_2)_S^r & (\theta_3)_S^r \end{vmatrix}, \ldots, \begin{vmatrix} (\theta_1)_1^R & (\theta_2)_1^R & (\theta_3)_1^R \\ (\theta_1)_2^R & (\theta_2)_2^R & (\theta_3)_2^R \\ \vdots & \vdots & \vdots \\ (\theta_1)_s^R & (\theta_2)_s^R & (\theta_3)_s^R \\ \vdots & \vdots & \vdots \\ (\theta_1)_S^R & (\theta_2)_S^R & (\theta_3)_S^R \end{vmatrix}$$

It can be shown (Geman and Geman, 1984) that $\{(\theta_1, \theta_2, \theta_3)_S^1, \ldots, (\theta_1, \theta_2, \theta_3)_S^r, \ldots, (\theta_1, \theta_2, \theta_3)_S^R\}$ is a sample of size R drawn from a c.d.f. F_S that converges, for large S, to the joint posterior distribution of $(\theta_1, \theta_2, \theta_3) \mid \mathbf{y}$. This sample will be used to estimate the posterior marginals. It may not be obvious to the reader that by drawing iteratively from conditionals we end up with a sample from the joint distribution. None the less, this is indeed what happens (see the aforementioned references for details).

The convergence of the algorithm is discussed in Gelfand and Smith (1990). Various techniques have been developed to check the convergence to the marginal distributions (see, for example, Zeger and Karim, 1991; Carlin *et al*, 1991). This remains an area of active research. Our experience suggests that R should be fairly large but that S need not be large. All the examples in this paper were done with $S = 15$. Typically, we found setting R to 200 or 400 to be sufficient. We judged this visually by repeating the entire process three times and then plotting the estimated marginals. For the last example of Section 5 we used $R = 1000$. Also, it is worth pointing out that the size of the resulting sample can be increased by including the last j values $\{(\theta_i)_{S-j+1}^r, (\theta_i)_{S-j+2}^r, \ldots, (\theta_i)_S^r\}$. This results in a final sample of size jR. Although the elements of the sample are not independent, the ergodic convergence of the process guarantees that we can still use this sample to estimate the posterior marginals. We did not use this technique in this paper.

Two methods can be used to estimate the posterior marginals from the samples. If the conditional posterior distributions are in closed form, then we estimate the density $f(\theta_1 \mid \mathbf{y})$, say, by $\hat{f}(\theta_1 \mid \mathbf{y})$ where

$$\begin{aligned} \hat{f}(\theta_1 \mid \mathbf{y}) &= \frac{1}{R} \sum_{r=1}^{R} f_1(\theta_1 \mid (\theta_2)_S^r, (\theta_3)_S^r, \mathbf{y}) \\ &\approx E_{\theta_2, \theta_3 \mid \mathbf{y}}(f_1(\theta_1 \mid \theta_2, \theta_3, \mathbf{y})) \\ &= \int\int f_1(\theta_1 \mid \theta_2, \theta_3, \mathbf{y}) f(\theta_2, \theta_3 \mid \mathbf{y}) \mathrm{d}\theta_2 \, \mathrm{d}\theta_3 = f(\theta_1 \mid \mathbf{y}) \end{aligned}$$

If, instead, the conditional distributions are not specified in a closed form, $\hat{f}(\theta_1 \mid \mathbf{y})$ is estimated from the sample $\{(\theta_1)_S^1, \ldots, (\theta_1)_S^R\}$ using a kernel estimator (Tapia and Thompson, 1978). An alternative to the kernel estimator is to use the above formula by evaluating the product of the likelihood and prior over a grid of values of θ_1 and normalizing this product by way of a one-dimensional integration. We are currently investigating this approach.

It is often the case that some of the parameters are conditionally independent. For example, it may turn out that $f(\theta_1 \mid \theta_2, \theta_3, \mathbf{y})$ does not depend on θ_2, say. This usually simplifies the procedure. Zeger and Karim (1989) discuss a simple graphical method for quickly identifying these instances of conditional independence.

3. The normal location-scale problem

We begin by considering the usual Gaussian location-scale problem. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a sample with density

$$f(y_i \mid \mu, \sigma^2, \epsilon, A_i) = (1 - \epsilon)\phi(y_i \mid \mu, \sigma^2) + \epsilon\phi(y_i \mid \mu + A_i, \sigma^2)$$

It is convenient to re-express the model as follows. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ be independent Bernoulli trials with success probability ϵ, and let $\mathbf{A} = (A_1, \ldots, A_n)$. Then,

$$y_i \mid \mu, \sigma^2, \mathbf{A}, \boldsymbol{\delta} \sim N(\mu + \delta_i A_i, \sigma^2)$$

Note that each y_i is conditionally independent of ϵ. First, consider the case where ϵ is known. We use the standard conjugate priors for μ and σ^2. That is, we assume that $\mu \sim N(\theta, v^2)$ and that σ^2 has an inverted χ^2 distribution with parameters ν and λ. We also consider the A_is to be independent, each with a $N(0, \tau^2)$ prior distribution.

To employ the Gibbs sampler we choose R arbitrary starting values for the $2 + 2n$ parameters μ_0^r, σ_0^r, $(\delta_i)_0^r, (A_i)_0^r$, $(i = 1, \ldots, n;\ r = 1, \ldots, R)$. To generate the random numbers, we need the densities $f_1(\mu \mid \mathbf{y}, \sigma^2, \boldsymbol{\delta}, \mathbf{A}, \epsilon)$,

$f_2(\sigma^2 | \mathbf{y}, \mu, \boldsymbol{\delta}, \mathbf{A}, \epsilon)$, $f_3(\boldsymbol{\delta} | \mathbf{y}, \mu, \sigma^2, \mathbf{A}, \epsilon)$ and $f_4(\mathbf{A} | \mathbf{y}, \mu, \sigma^2, \boldsymbol{\delta}, \epsilon)$ where it is understood that f_3 is a probability mass function. We now derive these conditional distributions.

First, note that conditional on the data and the other parameters, both μ and σ^2 are independent of ϵ. In order to obtain the distribution of $\mu | \sigma^2, \boldsymbol{\delta}, \mathbf{A}, \mathbf{y}$ we consider $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ where $y_i^* = y_i - \delta_i A_i$. Thus, $\mathbf{y}^*$ is identical to the original data except that the outliers have been corrected by subtracting off the location shift A_i. This correction is possible since both $\mathbf{A}$ and $\boldsymbol{\delta}$ are known at this step. Now, the y_i^*s are independent samples from a $N(\mu, \sigma^2)$ distribution, with σ^2 known. Thus, the formula for updating a conjugate prior (DeGroot, 1970) can be used and we have that $\mu | \mathbf{y}, \sigma^2, \boldsymbol{\delta}, \mathbf{A}$ is $N(a, b)$ where $a = \{\theta/v^2 + n\bar{y}^*/\sigma^2\}\{1/v^2 + n/\sigma^2\}^{-1}$, $b^{-1} = 1/v^2 + n/\sigma^2$, and $\bar{y}^*$ is the average of the y_i^*s. Standard routines can be used to generate the random draw of μ.

A similar argument is used for f_2. We find $\mathbf{y}^*$ and then employ the update formula for σ^2 with μ known. Thus, $\sigma^2 | \mathbf{y}, \mu, \boldsymbol{\delta}, \mathbf{A}$ has an inverted χ^2 distribution. Specifically, $(ns^2 + v\lambda)/\sigma^2$ has a χ^2 distribution with $n + v$ degrees of freedom, where $s^2 = \Sigma (y_i^* - \mu)^2/n$.

Now consider f_3. It is easy to see that, conditional on the data and the other parameters, each δ_i is an independent Bernoulli trial with success probability

$$p_i = \frac{\phi((y_i - \mu - A_i)/\sigma)\epsilon}{\phi((y_i - \mu - A_i)/\sigma)\epsilon + \phi((y_i - \mu)/\sigma)(1 - \epsilon)}$$

where $\phi(\cdot) \equiv \phi(\cdot | 0, 1)$ is the density function for a standard normal distribution.

Finally consider f_4. The A_is are conditionally independent given $\mu, \sigma, \epsilon, \boldsymbol{\delta}, \mathbf{y}$. We derive the conditional distribution for A_i. If $\delta_i = 1$, then $y_i - \mu$ is a sample from a $N(A_i, \sigma^2)$ distribution. Again, the standard conjugate update shows that $A_i | \mathbf{y}, \mu, \sigma^2, \delta_i = 1, \epsilon$ has a $N(c, d)$ distribution where

$$c = \frac{(y_i - \mu)/\sigma^2}{1/\tau^2 + 1/\sigma^2}$$

$$d^{-1} = 1/\tau^2 + 1/\sigma^2$$

If, instead, $\delta_i = 0$, then we have no information on A_i (that is, the likelihood for A_i is flat) so that the conditional distribution $A_i | \mathbf{y}, \mu, \sigma^2, \delta_i = 0, \epsilon$ is simply its prior distribution, namely, $N(0, \tau^2)$. If one wanted to use an improper prior for A_i, this could be approximated in the algorithm by using a large value for τ.

The Gibbs sampler can now be easily implemented. Standard routines can be used to generate random numbers from the required distributions. The procedure is repeated S times, resulting at the Sth step in samples

$$\mu_S^1, \ldots, \mu_S^R$$

$$(\sigma^2)_S^1, \ldots, (\sigma^2)_S^R$$

$$(\delta_i)_S^1, \ldots, (\delta_i)_S^R, \quad \text{for } i = 1, \ldots, n$$

$$(A_i)_S^1, \ldots, (A_i)_S^R, \quad \text{for } i = 1, \ldots, n$$

The posterior marginal for μ may be estimated as

$$\hat{f}(\mu | \mathbf{y}) = \frac{1}{R} \sum_{r=1}^{R} \phi(\mu | a_r, b_r)$$

where

$$a_r = \frac{\theta/v^2 + n\bar{y}^*/(\sigma^2)_S^r}{1/v^2 + n/(\sigma^2)_S^r}$$

$$b_r^{-1} = 1/v^2 + n/(\sigma^2)_S^r$$

$$y_i^* = y_i - (\delta_i)_S^r (A_i)_S^r$$

Similarly, the posterior probability that y_i is an outlier is estimated by

$$\frac{1}{R} \sum_{r=1}^{R} \frac{\epsilon\phi((y_i^* - \mu_S^r - (A_i)_S^r)/(\sigma)_S^r)}{\epsilon\phi((y_i^* - \mu_S^r - (A_i)_S^r)/(\sigma)_S^r) + (1 - \epsilon)\phi((y_i^* - \mu_S^r)/(\sigma)_S^r)}$$

The posterior for σ is of less interest and we do not bother to estimate it.

As an example, we used the infamous Darwin data (Fisher, 1960). The data consist of 15 height differences of cross- and self-fertilized plants. The two smallest observations (-67 and -48) are usually regarded as possible outliers. We used $\epsilon = 0.05$ and $v = \tau = 1000$. This value of ϵ is typically regarded as being reasonable for these data (Box and Tiao, 1968). The large values of v and τ are used to reproduce the usual flat priors used in this problem. Similarly, we take $v = 0$ to produce the non-informative prior for σ. But note that using informative priors does not make the calculations any more difficult. We used $R = 200$ and $S = 15$. For the purposes of illustration, we repeated our entire analysis three times to expose the variability of the procedure.

The estimated posteriors for μ and $\boldsymbol{\delta}$ are plotted in Figs 1a and 1b. When feasible, the plots include the three estimated posteriors corresponding to the three repeated runs. In some cases, such as Fig. 1b, it is difficult to distinguish the different runs by eye, so the results of only one run are displayed. An informal graphical inspection of the three densities in the plot suggests that the process has converged. Larger values of R can be used to reduce the variability even further. If the possibility of outliers is excluded by setting $\epsilon = 0$, the posterior for μ will have a mode at 20.933 (Box and Tiao, 1968). Allowing for outliers causes the posterior to be shifted to the right so that the effect of the two extreme observations has been downweighted (for similar analyses, see Box and Tiao, 1968; Abraham and Box, 1978). The two smallest observations are regarded as outliers in this data set and indeed, we see that the smallest observation ($y = -67$) has a posterior probability of almost 0.40 of being an outlier, while the second smallest observation ($y = -48$) has a posterior probability of slightly over 0.20 of being an outlier. The

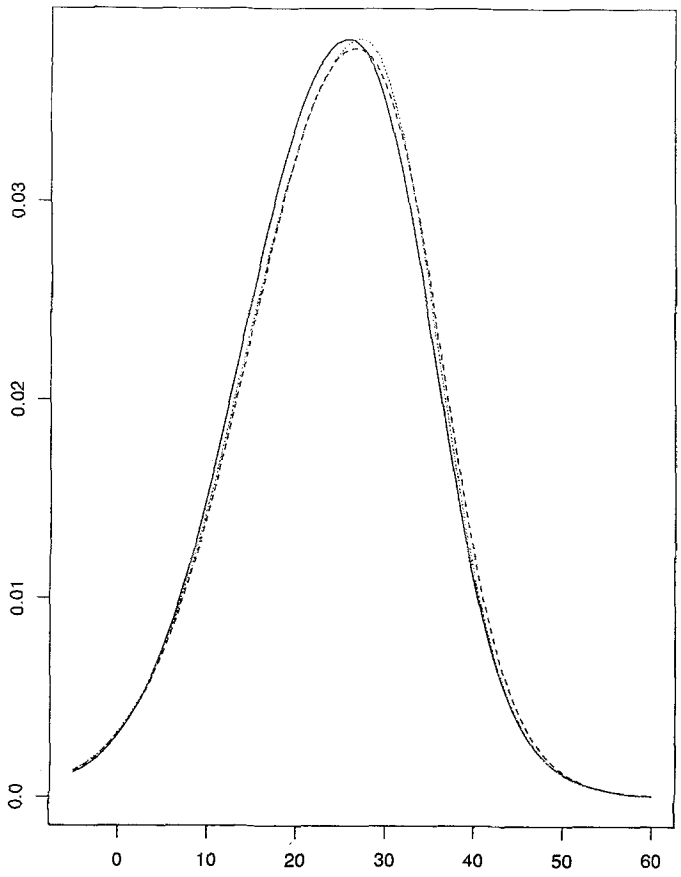


Fig. 1a. *Posterior marginals for μ based on a contaminated Normal model with $\epsilon = 0.05$ for the Darwin data. Each of the three curves corresponds to a complete run of the Gibbs sampler.*

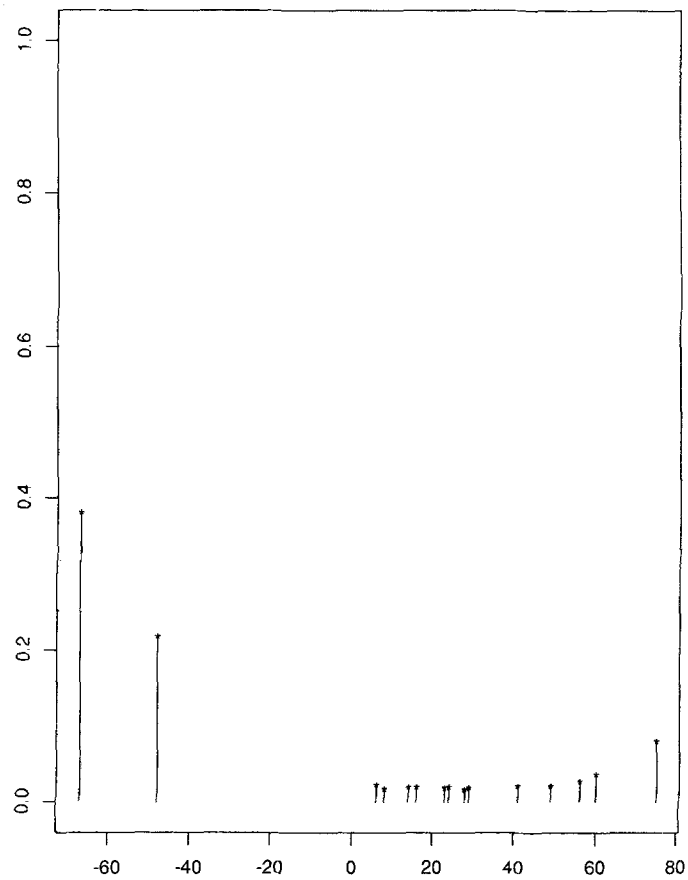


Fig. 1b. *Posterior marginal probability that each observation is an outlier based on a contaminated Normal model with $\epsilon = 0.05$ for the Darwin data.*

remainder of the observations have relatively small posterior probabilities of being outliers, with the largest being the last observation ($y = 75$), with probability 0.075. Note that the posterior probability that a subset of observations are outliers can be estimated by counting the proportion of the corresponding δ_is that were 1s on the sample from the Gibbs algorithm. In particular, to estimate the probability that there are no outliers, we need only count the proportion of times that all δ_is were 0s. This turned out to be 0.435. Also, the probability that there were one, two, three or four outliers is 0.335, 0.180, 0.035 and 0.015, respectively. The program to carry out these calculations is very simple and was implemented in New S (Becker *et al*, 1988).

We now consider the case where ϵ is unknown. To our knowledge, this case has not been treated before. We are thus adding another parameter to the model. Using the Gibbs sampler, this turns out to be extremely simple as well. The densities f_1, f_2, f_3 and f_4 remain unchanged since they are conditional on ϵ. But we need a fifth conditional posterior, namely, the distribution for ϵ conditional on the data and the rest of the parameters. We take ϵ to have a beta(ρ_1, ρ_2) prior distribution. A flat prior is not appropriate, for, presumably, we would not be conducting the experiment if there were an exceedingly large probability of many outliers. As mentioned before, several authors have considered a value of 0.05 to be reasonable for ϵ (Box and Tiao, 1968). We thus take the mean of the prior for ϵ to be 0.05. It seems reasonable to assume that an observation has less than half a chance of being an outlier with high probability. If we assume that $P(\epsilon < 0.5) = 0.99$, then this implies that $\rho_1 = 0.1842$ and $\rho_2 = 3.5$. Now, since we are conditioning on $\boldsymbol{\delta}$, this posterior is straightforward. It is readily seen that the conditional distribution for ϵ depends only on $\boldsymbol{\delta}$. Let k be the number of δ_is that are equal to one. Then we simply have a binomial experiment with k successes and a beta prior. Hence the required density is beta$(\rho_1 + k, \rho_2 + n - k)$. Again, we can generate random numbers from this distribution using standard routines. The posterior marginal for ϵ is estimated by

$$\hat{f}(\epsilon \mid \mathbf{y}) = \frac{1}{R} \sum_{r=1}^{R} B(\rho_1 + k(r), \rho_2 + n - k(r))$$

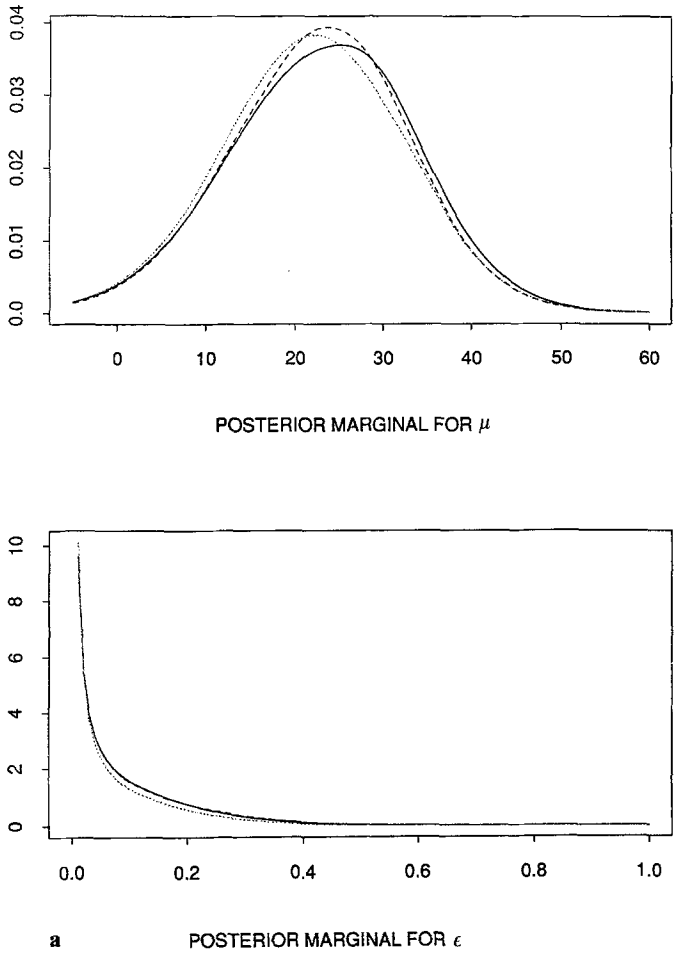


Fig. 2a. *Posterior marginal for μ and ϵ based on a contaminated Normal model with ϵ unknown for the Darwin data.*

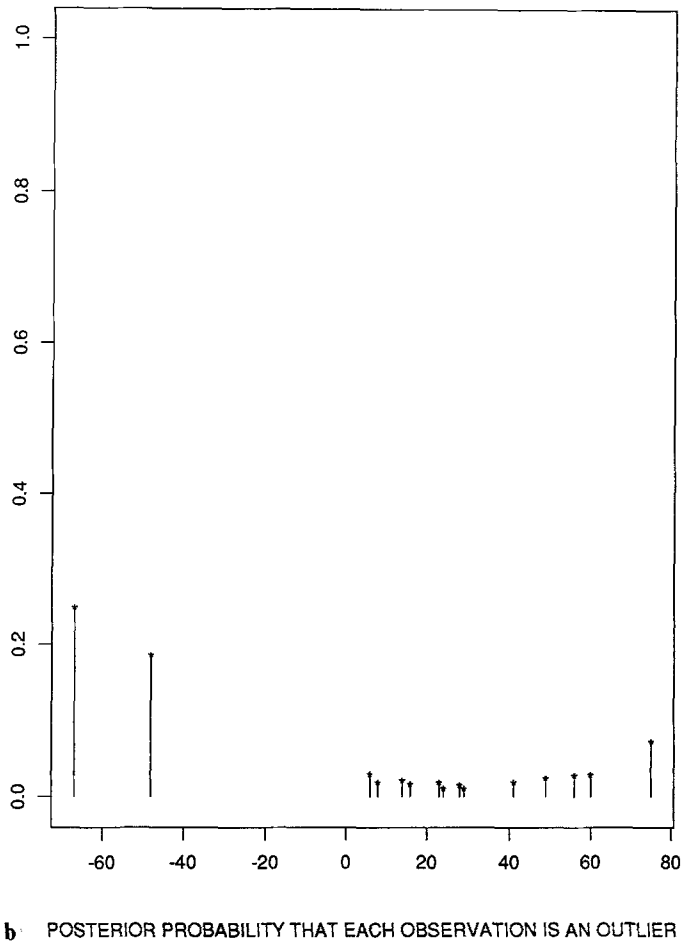


Fig. 2b. *Posterior marginal probability that each observation is an outlier based on a contaminated Normal model with ϵ unknown for the Darwin data.*

where $B(\rho_1, \rho_2)$ is the density for a beta(ρ_1, ρ_2) distribution and $k(r)$ is the number of the $\{(\delta_1)'_S, \ldots, (\delta_n)'_S\}$ that are equal to one.

We repeated the analysis of the Darwin data using this prior for ϵ. The resulting estimated marginals for μ, δ and ϵ are shown in Figs 2a and 2b. The posterior for μ is essentially the same as in the case where ϵ is known. However, the posterior probability that the first observation is an outlier has dropped below 0.4. This suggests that our uncertainty about ϵ lowers our certainty that this observation is an outlier. Specifically, since small values of ϵ are possible, our posterior probability that the observation is an outlier is lowered, though this still stands out as a likely outlier relative to the other observations. More importantly, allowing for the possibility of outliers makes the inferences for μ robust. Although the actual probability attached to a particular observation being an outlier is affected by the prior for ϵ, the inferences for μ are stable, once the possibility of outliers is included in some way. We emphasize that adding in the extra unknown was extremely simple and involved adding one extra subroutine to the algorithm to draw random numbers from f_5. Thus, increasing the dimension of the problem requires only simple changes to the algorithm. Finally, we estimated the probabilities of there being zero, one, two, three or four outliers to be 0.670, 0.110, 0.095, 0.060, and 0.045, respectively.

We point out that going from the ϵ known case to ϵ unknown, is like switching from a prior on ϵ that is a point mass at one value, to a smooth prior. Although we saw that the analysis changes, it does not change drastically. This gives us some confidence that the analysis is not too dependent on the choice of prior for ϵ.

4. The t-distribution

The Student t-distribution has been proposed as an alternative sampling model when extreme observations are considered a possibility. Estimates of the location parameter μ are then less affected by extreme observations. We show in this section that the Gibbs sampler can be applied

for finding the marginal posterior for μ in this case. Thus, we assume that the density for y_i is

$$f(y_i \mid \mu, \sigma^2, \alpha) \propto \left[\left(\frac{y_i - \mu}{\sigma}\right)^2 + \alpha\right]^{-\frac{\alpha+1}{2}}$$

It is convenient to introduce, for each observation, an extra parameter W_i such that

$$y_i \mid \mu, \sigma^2, W_i \sim N(\mu, \sigma^2 W_i^{-1})$$

and to assume that the W_is are independently distributed as gamma random variables with parameters $(\alpha/2, \alpha/2)$. It is straightforward to see that the marginal distribution for each observation

$$f(y_i \mid \mu, \sigma^2, \alpha) = \int f(y_i \mid \mu, \sigma^2, W_i) f(W_i \mid \alpha) \mathrm{d}W_i$$

is then a t-distribution with α degrees of freedom. In this way, we have simply rewritten the t-distribution as a mixture of normals introducing n extra parameters W_i. Note that each y_i is conditionally independent of α given W_i. Distributions other than the Student t can be expressed as a mixture of normals (see, for example, West, 1987). Hence, it is trivial to adapt the methods in this section to deal with any scale mixture of normals simply by replacing the gamma distribution with the required mixing distribution. The Gibbs sampler can thus be used to develop a complete Bayesian analysis for a large class of models.

The set of parameters that we are going to consider here is $\{\mu, \sigma^2, \mathbf{W}, \alpha\}$ where $\mathbf{W} = (W_1, W_2, \ldots, W_n)$. Let us assume, at first, that α is known and let the prior distributions for μ and σ^2 be, as in Section 3, normal and inverted chi-squared. Values for α in the range of 1 to 7 have been cited as being reasonable (see Fraser, 1979, p. 37; Lange *et al*, 1989). For now, we shall take $\alpha = 3$. The conditional posterior distributions required to implement the Gibbs algorithm can be readily derived. That is, it can be seen through successive application of the conjugate update formulas, together with the fact that $y_i \mid \mu, \sigma^2, W_i$ has a $N(\mu, \sigma^2 W_i^{-1})$ distribution, that $\mu \mid \sigma^2, \mathbf{W}, \mathbf{y} \sim N(a, b)$, where

$$a = \frac{\theta/v^2 + (\Sigma\, y_i W_i)/\sigma^2}{v^{-2} + (\Sigma\, W_i)\sigma^{-2}}$$

$$b^{-1} = \frac{1}{v^2} + \frac{\Sigma\, W_i}{\sigma^2}$$

Further, we obtain that $\sigma^2 \mid \mu, \mathbf{W}, \mathbf{y}$ has an inverted chi-squared distribution. More precisely:

$$\frac{\Sigma\, W_i (y_i - \mu)^2 + v\lambda}{\sigma^2} \sim \chi^2_{n+v}$$

Also, we have the W_is are conditionally independent and $W_i \mid \mu, \sigma^2, \mathbf{y}$ is a gamma distribution with parameters $(\alpha + 1)/2$ and $[(y_i - \mu)^2/\sigma^2 + \alpha]/2$.

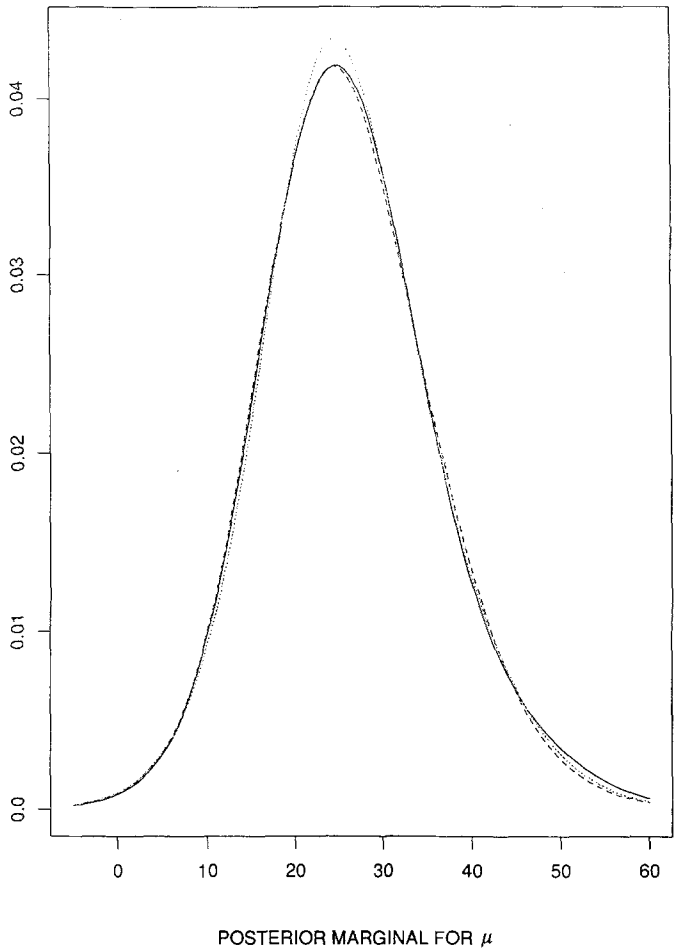


Fig. 3. *Posterior marginal for μ based on a t distribution with three degrees of freedom for the Darwin data.*

Therefore, all the required conditional posterior distributions are available analytically and, as described in Section 2, we proceed generating random numbers from these distributions using standard routines. After S iterations of the algorithm we obtain samples of size R for the $n + 2$ parameters considered.

In particular, the marginal posterior density for μ for the Darwin data, shown in Fig. 3, has been estimated from the samples $(\sigma^2)^1_S, \ldots, (\sigma^2)^r_S, \ldots, (\sigma^2)^R_S$ and $(W_i)^1_S, \ldots, (W_i)^r_S, \ldots, (W_i)^R_S$, for $i = 1, 2, \ldots, n$ by

$$\hat{f}(\mu \mid \mathbf{y}) = \frac{1}{R}\sum_{r=1}^{R} f(\mu \mid (\sigma^2)^r_S, (\mathbf{W})^r_S, \mathbf{y}) = \frac{1}{R}\sum_{r=1}^{R} \phi(\mu \mid a_r, b_r)$$

where

$$a_r = \frac{\theta/v^2 + (\Sigma\, y_i (W_i)^r_S)/(\sigma^2)^r_S}{v^{-2} + (\Sigma\, (W_i)^r_S)(\sigma^{-2})^r_S}$$

$$b_r^{-1} = \frac{1}{v^2} + \frac{\Sigma\, (W_i)^r_S}{(\sigma^2)^r_S}$$

We found that the estimates based on $R = 200$ were not as

stable as in the contaminated normal case. The estimates in Fig. 3 are based on $R = 400$.

So far, the shape parameter α has been considered as known. It is far more realistic to assume that α is unknown. We now proceed with α regarded as an unknown parameter.

As α varies from 1 to ∞ the class of t-distributions varies from the Cauchy to the normal. It is convenient to define $\beta = 1/\alpha$ and we use a beta distribution for β with parameters 1.75 and 2.5. In this way, the mode of β is 1/3. Also we have that $P\{1/3 < \beta < 1/2\} = 0.26$, $P\{\beta > 1/3\} = 0.6$, $P\{\beta < 1/10\} = 0.06$. We tried to choose a prior that reflected a fair amount of uncertainty about β, that had a mode at a reasonable value ($\beta = 1/3$) and not too much probability near normality. We do not claim that this is an optimum prior in any sense. An interesting topic for futher research is to determine reasonable prior distributions for parameters that index outlier models. One referee suggested putting a point mass on the normal model ($\beta = 0$). This would especially be appropriate if one were interested in computing a Bayes factor for the hypothesis that the normal model is correct. However, since our goal is to protect us from outliers by allowing for the possibility of extreme observations, rather than model identification, we feel that the prior we have chosen is sufficient.

This is a case in which the conditional posterior distribution of β is not in closed form. In fact, a prior for β, or α, cannot be expressed within a conjugate family; hence we only have the kernel of the distribution:

$$f(\beta \mid \mathbf{W}, \mu, \sigma, \mathbf{y}) = f(\beta \mid \mathbf{W}) \propto f(\beta) \prod_{i=1}^{n} f(W_i \mid \beta)$$

$$\propto \frac{\beta^{1.75-1}(1-\beta)^{2.5-1}}{[(2\beta)^{1/2\beta}\Gamma(1/2\beta)]^n} \left\{\prod_{i=1}^{n} W_i\right\}^{1/2\beta - 1} \exp\left\{-\frac{1}{2\beta}\sum_{i=1}^{n} W_i\right\}$$

Here, $f(\beta)$ is the prior density for β. Note that we are dealing here with the distribution of β conditional on $\mathbf{W}$, since the introduction of the new variables W_is makes the model for the data independent of the shape parameter $1/\beta$. In other words, β is only influenced by $\mathbf{W}$ and the effect of the data on β is through $\mathbf{W}$.

Random numbers can be drawn from $f(\beta \mid \mathbf{W})$ using the rejection method (Section 2). Then the marginal posterior

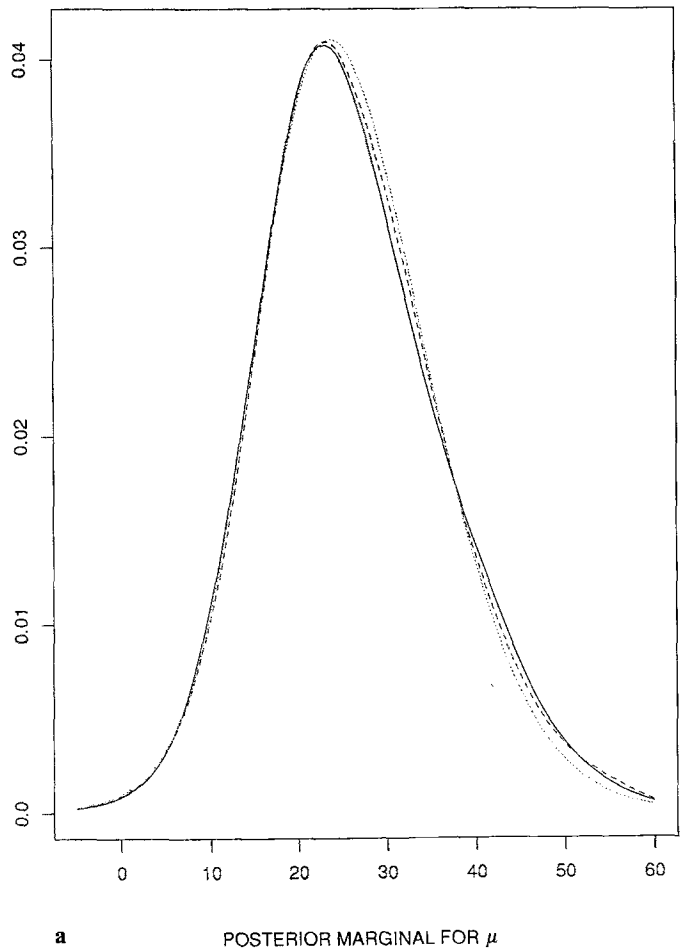


Fig. 4a. *Posterior marginal for μ based on a t distribution with degrees of freedom α unknown for the Darwin data.*

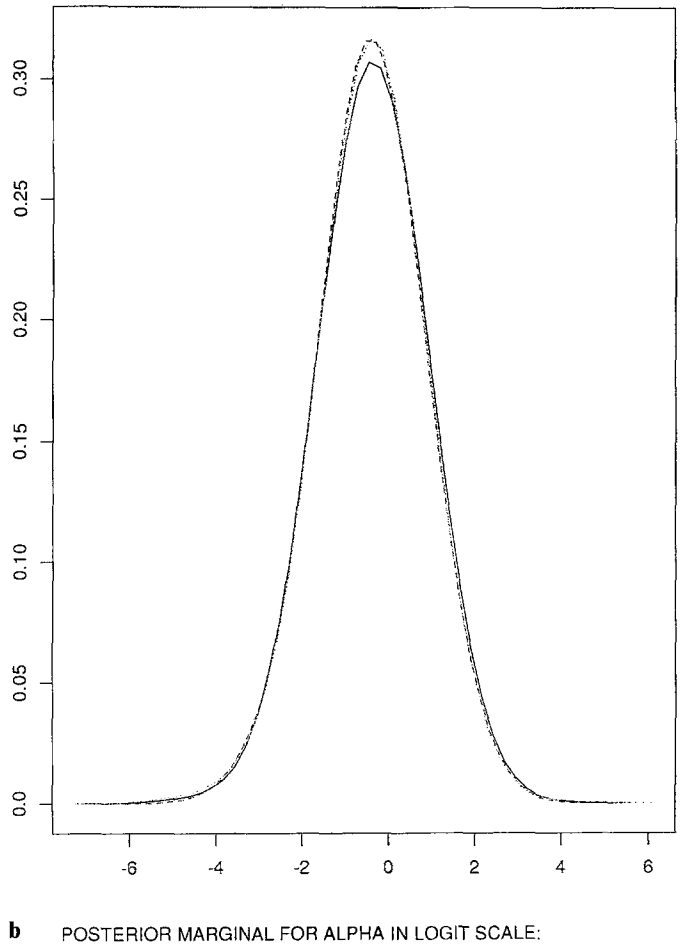


Fig. 4b. *Posterior marginal for logit(β) based on a t distribution with degrees of freedom α unknown for the Darwin data where $\alpha = 1/\beta$.*

distribution can be estimated from the sample by a kernel density estimator. The marginal for μ is plotted in Fig. 4a using $R = 400$. We see that the marginal for μ is not greatly affected by our uncertainty about β. We have also obtained the posterior for β using a kernel density estimator; see Fig. 4b. As expected, this posterior is very similar to the prior since it requires a large amount of data to learn about tail thickness. The reason for including β as an unknown parameter is not to learn about β, but rather to lead to a more honest analysis for μ.

It is interesting to consider a different parameterization for the shape parameter. Let $\gamma = \log(\beta/(1-\beta))$ be the logit of β. When $\gamma \to -\infty$ the sampling distribution is normal and when $\gamma \to +\infty$ the sampling distribution is Cauchy. The origin corresponds to $\alpha = 2$. This suggests that values usually employed, namely around $\alpha = 2$ and $\alpha = 3$ are, in some sense, mid-way between the normal and the Cauchy distributions.

5. Binomial models

Winkler and Gaba (1990) considered the following problem. We sample $\mathbf{w} = (w_1, \ldots, w_n)$ where each w_i is an independent Bernoulli trial with success probability p. Then, each w_i is either switched or not switched, where the switching takes place with probability ϵ. That is, we observe y_i where, conditional on w_i, $y_i = (1-\delta_i)w_i + \delta_i(1-w_i)$ and each δ_i is a Bernoulli trial with success probability ϵ. We let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$. This may be viewed as a binomial version of the outlier problem. In this case, an outlier is simply a Bernoulli observation that has been switched. The posterior marginals for p and ϵ are derived in Winkler and Gaba (1990). As in the Gaussian case, the computations are not simple. We point out that p and ϵ are not identifiable in this model. We will be using proper priors, however, so that the posterior marginals are well defined. In this situation, the prior information is quite important.

To use the Gibbs sampler, we need $f(p \mid \epsilon, \mathbf{y}, \boldsymbol{\delta})$, $f(\epsilon \mid p, \mathbf{y}, \boldsymbol{\delta})$ and $f(\delta_i \mid p, \epsilon, \mathbf{y})$. Following Winkler and Gaba (1990) we use a beta(α, β) prior for p and a beta(a, b) prior for ϵ. We now derive the three required conditional posterior distributions.

Consider $p \mid \epsilon, \mathbf{y}, \boldsymbol{\delta}$. Since $\boldsymbol{\delta}$ is given, we can define the corrected data vector $\mathbf{w} = (w_1, \ldots, w_n)$ where $w_i = (1-\delta_i)y_i + \delta_i(1-y_i)$. Then, $\mathbf{w}$ is a vector of Bernoulli observations so that the required conditional posterior is beta$(\alpha + \Sigma w_i, \beta + n - \Sigma w_i)$. Similarly, $\epsilon \mid p, \mathbf{y}, \boldsymbol{\delta}$ has a beta$(a + \Sigma \delta_i, b + n - \Sigma \delta_i)$ distribution. Finally, it is easy to see that $\delta_i \mid p, \epsilon, \mathbf{y}$ is Bernoulli with success probability q_i where

$$q_i = \frac{\epsilon p^{w_i}(1-p)^{1-w_i}}{\epsilon p^{w_i}(1-p)^{1-w_i} + (1-\epsilon)p^{y_i}(1-p)^{1-y_i}}$$

It is straightforward to generate the required random numbers. After S iterations, this leads to samples

$$p_S^1, \ldots, p_S^R$$
$$\epsilon_S^1, \ldots, \epsilon_S^R$$
$$(\delta_i)_S^1, \ldots, (\delta_i)_S^R, \quad i = 1, \ldots, n$$

The estimates of the posterior marginals are thus

$$\hat{f}(p \mid \mathbf{y}) = \frac{1}{R}\sum_{r=1}^{R} B\left(\alpha + \sum_i (w_i)_S^r, \beta + n - \sum_i (w_i)_S^r\right),$$

$$\hat{f}(\epsilon \mid \mathbf{y}) = \frac{1}{R}\sum_{r=1}^{R} B\left(a + \sum_i (\delta_i)_S^r, b + n - \sum_i (\delta_i)_S^r\right)$$

and

$$\hat{P}(\delta_i = 1 \mid \mathbf{y}) = \frac{1}{R}\sum_{r=1}^{R} \frac{\epsilon (p_S^r)^{(w_i)_S}(1-p_S^r)^{1-(w_i)_S}}{\epsilon (p_S^r)^{(w_i)_S}(1-p_S^r)^{1-(w_i)_S} + (1-\epsilon)(p_S^r)^{y_i}(1-p_S^r)^{1-y_i}}$$

We applied this to a data set involving self reported delinquent behaviour (Gould, 1969) as analysed by Winkler and Gaba. Of 104 college students who were asked if they had beaten someone up, 21 said they had. Here, $y_i = 1$ corresponds to an affirmative response. Based on information from Clark and Tifft (1966), Winkler and Gaba proposed a beta(2, 8) distribution for p and a beta(2, 18) distribution for ϵ. The resulting posteriors for p and ϵ, based on $R = 400$, are shown in Fig. 5. The posteriors are the same as those displayed in Winkler and Gaba (1990). The probability that y_i is an outlier is 0.35 if $y_i = 1$ and is 0.01 if $y_i = 0$. It could be agreed that the outlier model should be expanded to allow a probability ϵ_1 of a 1 being switched to a 0 and a probability ϵ_2 of a 0 being switched to a 1. It is obvious how to generalize the Gibbs sampler to deal with this case; the calculations are not any more difficult. We shall not pursue these details here.

We now consider the more interesting case where a regressor x is available. Specifically, suppose that each w_i is Bernoulli with success probability

$$p_i(\alpha, \beta) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

where α and β are unknown regression parameters and x_i is the observed value of some predictor variable x. This is the standard logistic regression model (McCullagh and Nelder, 1989, p. 108).

The problem of outliers in logistic regression is dealt with in Pregibon (1981; 1982) and Copas (1988). In his discussion of Copas, O'Hagan (1988) suggests that a Bayesian approach is possible, though he acknowledges the heavy computational burden that this entails. And Davison (1988), discussing the same paper, shows that Laplace's method of approximating integrals can be used to approximate the predictive probability of an observa-

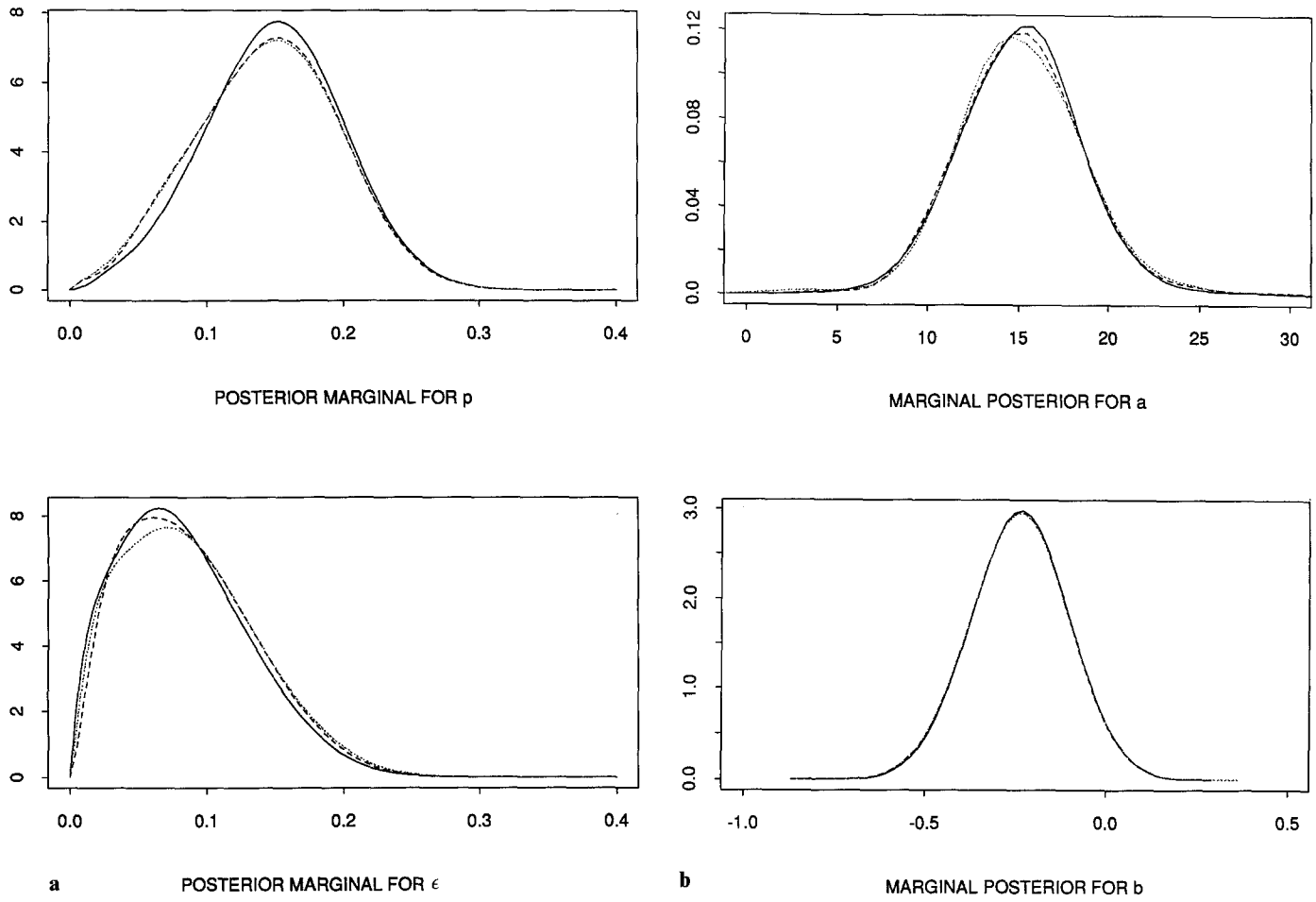


Fig. 5. *Posterior marginals for p and ϵ for the Binomial model.*

tion given the rest of the data and thus he obtains a Bayesian method for identifying suspicious observations. Here, as in the previous cases, we consider a completely Bayesian approach. As before, we assume that there is a probability ϵ that w_i is switched to $1 - w_i$. More formally, we assume that $y_i = (1 - \delta_i)w_i + \delta_i(1 - w_i)$ where δ_i is Bernoulli with success probability ϵ. This is the model used by Copas (1988). Ekholm and Palmgren (1982) and Palmgren and Ekholm (1987) propose more general models for contaminated binary data. We shall not pursue those models here, though it should be possible to extend our methods to deal with those models.

The likelihood function based on the uncontaminated data $\mathbf{w}$ is

$$L_{\mathbf{w}}(\alpha, \beta, \epsilon) = \prod_i p_i(\alpha, \beta)^{w_i}(1 - p_i(\alpha, \beta))^{1 - w_i}$$

Convenient conjugate priors do not exist for this model. This does not present a serious problem for our analysis. To sample from the conditional distributions we resort to the rejection method described in Section 2. We now describe the process in more detail. We took ϵ to be a priori independent of α and β with a $B(a, b)$ distribution. We let α and β have an arbitrary prior density denoted by $\pi(\alpha, \beta)$.

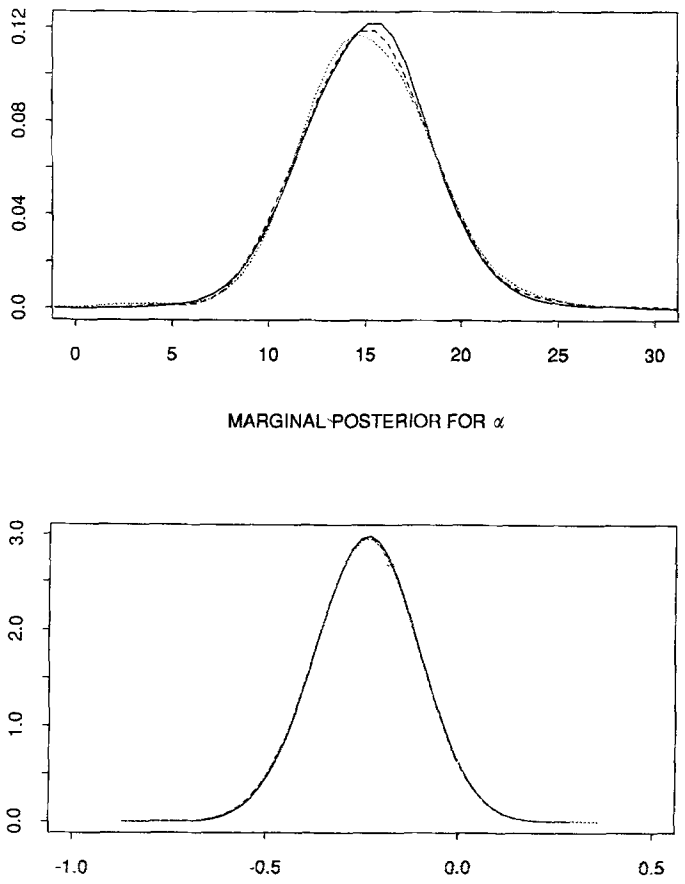
First we consider $f(\alpha \mid \beta, \epsilon, \boldsymbol{\delta}, \mathbf{y}, x)$. Let $w_i = (1 - \delta_i)y_i + \delta_i(1 - y_i)$. The posterior conditional for α is proportional to $L_{\mathbf{w}}(\alpha, \beta, \epsilon)\pi(\alpha, \beta, \epsilon)$ where β and ϵ are fixed. We draw a random α from this distribution using the rejection method described in Section 2. The method for drawing β is the same.

We can find $f(\epsilon \mid \alpha, \beta, \boldsymbol{\delta}, \mathbf{y}, x)$ in closed form. First note that ϵ is conditionally independent of α, β, $\mathbf{y}$ and x. Let $k = \Sigma\, \delta_i$. Then obviously, the conditional distribution for ϵ is $B(a + k, b + n - k)$.

Finally, consider $\boldsymbol{\delta} \mid \alpha, \beta, \epsilon, \mathbf{y}, x$. Each δ_i is Bernoulli with success probability

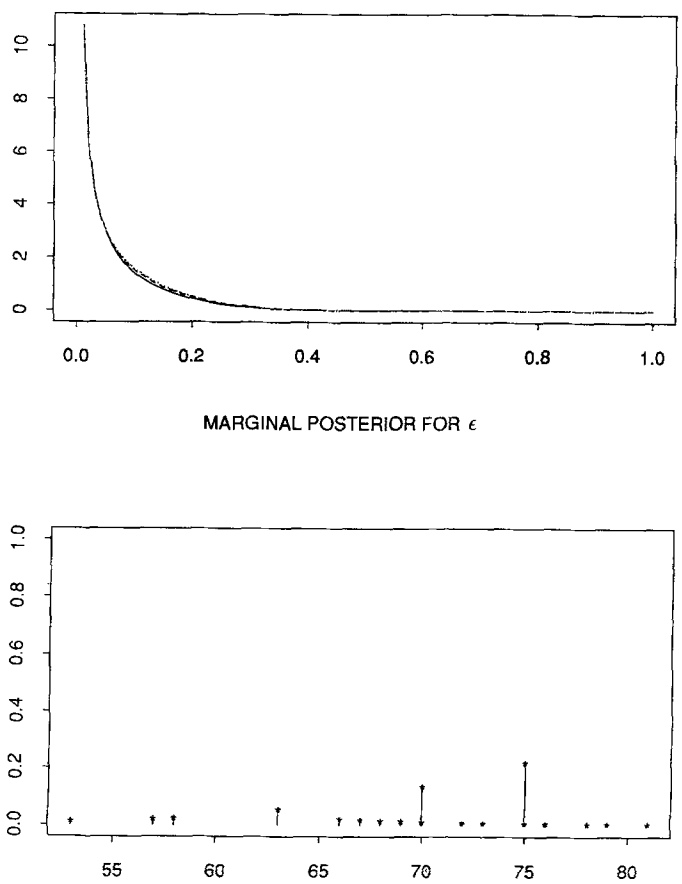
$$\frac{\epsilon p_i^{1 - y_i}(1 - p_i)^{y_i}}{\epsilon p_i^{1 - y_i}(1 - p_i)^{y_i} + (1 - \epsilon)p_i^{y_i}(1 - p_i)^{1 - y_i}}$$

where $p_i \equiv p_i(\alpha, \beta)$.



a

Fig. 6a. *Posterior marginals for α and β for the Challenger data.*



b

Fig. 6b. *Posterior marginals for ε and posterior marginal probability that each observation is an outlier for the Challenger data.*

We illustrate the method with data from Dalal *et al* (1989). They analysed data on 23 launches of the space shuttle to estimate the probability of an O-ring failure on the day the space shuttle Challenger was launched. The data we used was a binary outcome, indicating an incident with the O-rings (erosion or blowby) and the regressor was joint temperature. Following Dalal *et al*, we used a logistic model but we allowed for outliers. We used uniform priors for α and β and a beta(0.1842, 3.5) prior for ϵ as in Section 3. Posterior marginals were estimated from the samples by kernel density estimation, as suggested in Gelfand and Smith (1990). Because we relied on kernel density estimators, we increased R to 1000 but kept $S = 15$. We used FORTRAN since S was too slow in this case.

The estimated posteriors are in Figs. 6a and 6b. The O-ring failure at 75°F is an outlier with probability 0.22. This is consistent with the Dalal *et al* analysis. Similarly, the failures at 70°F each have probability 0.14 of being outliers. The failure of 63°F has probability 0.05 of being an outlier. While this is not very large, it is interesting to note that the Dalal *et al* analysis does not seem to flag the observations at 63°F at all. This may be a masking effect. As O'Hagan (1988) remarks, masking is not a problem in the Bayesian approach since we are computing the marginal probability that each observation is an outlier. This averages over the possibilities that there are simultaneously other outliers.

6. Discussion

Our emphasis has been on demonstrating the simplicity of the Gibbs sampling approach to the computation of posterior marginals in outlier problems. Other approaches may well provide faster, more efficient routines for doing the same task, but what the Gibbs approach has to offer is greater flexibility and less programming effort. That is, less computer efficiency is traded for more human efficiency.

The conceptual simplicity of the approach makes it easy to write the necessary programs and, having written the programs, it is easy to change them to handle new

problems. For example, to go from ϵ known to ϵ unknown in the ϵ contamination model, we needed only add one new subroutine, namely a routine to draw from the conditional distribution for ϵ. In S, this involved adding one new function of four lines of code. Similarly, to replace the t-distribution with a different mixture of normals would involve replacing the gamma function with the appropriate mixing distribution. Thus, one small part of the program would be modified: instead of drawing random gammas we would draw from a different distribution.

The methods in Section 3 and 4 can easily be extended to the regression case. Instead of drawing μ from a normal distribution, for example, we would draw $\boldsymbol{\beta}$ from a multivariable normal distribution, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are the regression parameters. Similarly, the other conditional distributions would be adapted in the obvious way. In the case of logistic regression, sampling from the posterior conditionals of the regressors $\beta_1, \ldots, \beta_p$ would best be done one at a time. But the sampling method would be exactly the same as that used in Section 5 so we would essentially just repeat that part of the program p times. The point is that increasing the dimension of the problem increases the execution time of the program but it does not increase the complexity of the program. There is no need to deal with high-dimensional grids, for example. In a sense, the Gibbs sampler replaces a difficult high-dimensional problem with a series of simple one-dimensional problems.

It is easy to adapt the Gibbs sampler to deal with outliers for other sampling models. For example, Pettit (1988) considers sampling from an exponential when outliers are possible. He derives an insightful approximation to find the Bayes factor for a particular observation being an outlier. Using the Gibbs sampler for this problem is essentially the same as the contaminated normal case in Section 3. In particular, at most steps in the algorithm, we are conditioning on the vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ that tells us which observations are outliers so that the outliers can be corrected and the usual Bayesian conjugate distributions can be used. And the distribution of $\boldsymbol{\delta}$ given the other parameters is straightforward. Thus, a fully Bayesian analysis is possible and the posterior probability that a particular observation is an outlier can be estimated, as can the posterior probability that a set of observations is an outlier.

Acknowledgments

We are grateful to Rob Kass for suggesting that we apply these methods to the t-distribution and for many other useful discussions. We also thank Brad Carlin, Nick Polson, Joel Greenhouse and two referees for helpful comments. This work was completed while the first author was visiting the Department of Statistics at Carnegie Mellon University and was partially supported by the Italian Research Council (CNR). The second author was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Abraham, B. and Box, G. E. P. (1978) Linear models and spurious observations. *Applied Statistics*, **27**, 131–138.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*, Wadsworth, Pacific Grove, California.

Box, G. E. P. and Tiao, G. C. (1968) A Bayesian approach to some outlier problems. *Biometrika*, **55**, 119–129.

Carlin, B., Gelfand, A. E. and Smith, A. F. M. (1989) Hierarchical Bayes analysis of change point problems. To appear in *Applied Statistics*.

Chaloner, K. and Brant, R. (1988) A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651–660.

Clark, J. P. and Tifft, L. L. (1966) Polygraph and interview validation of self reported deviant behavior. *American Sociological Review*, **31**, 516–523.

Copas, J. B. (1988) Binary regression models for contaminated models. *Journal of the Royal Statistical Society Series B*, **50**, 225–265.

Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1969) Risk analysis of the space shuttle: pre-Challenger prediction of failure. *Journal of the American Statistical Association*, **84**, 945–957.

Davison, A. C. (1988) Discussion on Copas. *Journal of the Royal Statistical Society Series B*, **50**, 258–259.

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

DeVroye, L. (1986) *Non-uniform Random Variate Generation*. Springer-Verlag, New York.

Ekholm, A. and Palmgren, J. (1982) A model for a binary response with misclassification. *GLIM-82: Proceedings of the International Conference on Generalized Linear Models*, R. Gilchrist (Ed), Springer-Verlag, Heidelberg, pp. 128–43.

Fisher, R. A. (1960) *The Design of Experiments* (7th edn), Oliver and Boyd, Edinburgh.

Fraser, D. A. S. (1979) *Inference and Linear Models*, McGraw-Hill, New York.

Freeman, P. R. (1980) On the number of outliers in data from a linear model, in *Bayesian Statistics*, Proceedings of the First International Meeting held in Valencia (Spain), J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (eds), 347–382.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*.

German, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Gould, L. L. (1969) Who defines delinquency: A comparison of self-reported and officially reported indices of delinquency for three racial groups. *Social Problems*, **16**, 325–336.

Guttman, I., Dutter, R. and Freeman, P. (1978) Care and handling of univariate outliers in the general linear model to detect spuriosity—a Bayesian approach. *Technometrics*, **20**, 187–194.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, **84**, 881–896.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models* (2nd edn), Chapman and Hall, London.

O'Hagan, A. (1988) Discussion on Copas. *Journal of the Royal Statistical Society Series B*, **50**, 259.

Palmgren, J. and Ekholm, A. (1987) Exponential family nonlinear models for categorical data with error of observation. *Appl. Stochast. Model Data Anal.*, **3**, 111–124.

Pettit, L. I. (1988) Bayes methods for outliers in exponential samples. *Journal of the Royal Statistical Society Series B*, **50**, 371–380.

Pettit, L. I. and Smith, A. F. M. (1984) *Bulletin of the International Statistical Institute: Proceedings of the 44th session*, 292–306.

Pettit, L. I. and Smith, A. F. M. (1985) Outliers and influential observations in linear models, in *Bayesian Statistics 2*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (eds), North-Holland, Amsterdam, pp. 473–494.

Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.* **9**, 705–724.

Pregibon, D. (1982) Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, **38**, 485–498.

Tanner, M. and Wong, W. (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–550.

Tapia, R. A. and Thompson, J. R. (1978) *Nonparametric Probability Density Estimation*, Johns Hopkins University Press, Baltimore, Maryland.

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

West, M. (1984) Outlier models and prior distributions in Bayesian linear regression. *J. R. Statist. B.*, **46**, 431–439.

West, M. (1987) On scale mixtures of normal distributions. *Biometrika*, **74**, 646–648.

Winkler, R. L. and Gaba, A. (1990) Inference with imperfect sampling from a Bernoulli process, in *Bayesian and Likelihood Methods in Statistics and Econometrics*, S. Geisser, J. S. Hodges, S. J. Press and A. Zellner (eds), North-Holland, Amsterdam.

Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.