Working Paper 95-21
Statistics and Econometrics Series 5
June 1995

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe, Madrid (Spain)
Fax (341) 624-9849

GIBBS SAMPLING WILL FAIL IN OUTLIER PROBLEMS WITH STRONG MASKING

Ana Justel and Daniel Peña [1]

Abstract

This paper discusses the convergence of the Gibbs sampling algorithm when it is applied to the problem of outlier detection in regression models. Given any vector of initial conditions, theoretically, the algorithm converges to the true posterior distribution. However, the speed of convergence may slow down in a high dimensional parameter space where the parameters are highly correlated. We show that the effect of the leverage in regression models makes very difficult the convergence of the Gibbs sampling algorithm in sets of data with strong masking. The problem is illustrated in several examples.

Key words:
Bayesian analysis, leverage, linear regression, scale contamination.

[1] Departament of Statistics and Econometrics, Universidad Carlos III de Madrid.

1 INTRODUCTION

The intensive attention that Gibbs sampling (Geman and Geman, 1984 and Gelfand and Smith, 1990) has received in applied work is due to its mild implementation requirements together with its programming simplicity. In a Bayesian parametric model this algorithm provides an accurate estimation of the marginal posterior densities, or summaries of these distributions, by sampling from the conditional parameter distributions. Furthermore, the algorithm converges independently of the initial condition and, in many applications, in a few iterations. However, several authors have indicated problems of convergence with Gibbs sampling. Gelman and Rubin (1992) showed the importance of the initial conditions in the speed of convergence of the algorithm in a high dimensional parameter problem. Matthews (1993) gave an example in which the Gibbs sampler seemed to converge when in fact it had not. Hills and Smith (1992) stressed that the number of iterations to achieve convergence is a function of the starting values and the correlation structure of the stochastic process generated by the Gibbs sampling. They concluded that the higher the correlation the more serious the convergency problem. Polson (1994) analysed a convergence rate bound that can be used to choose the number of iterations to guarantee desired sampling accuracy. The running times depends on the effects of correlation and dimension. Smith and Roberts (1993) and Mengersen and Robert (1994) pointed out that when the parameter distribution is bimodal, the Gibbs sampling iterations may be trapped in one of the modes, reducing the probability of reaching convergence.

In this paper we show that in the linear regression set up outliers can make very unlikely the convergence when there is a strong masking. If there are outliers which mask or swamp other observations, the parameter structure will be highly correlated and convergence will usually not be reached in a reasonable amount of iterations. In addition, the algorithm may provide a false idea of the posterior probabilities. In summary, in data set with masked high leverage outliers, the Gibbs sampling iterations are stable around wrong limit values for thousands of iterations.

This paper is organized as follows. Section 2 presents the Gibbs sampling application to detect outliers in linear regression problems by using the scale contaminated regression model and examines the algorithm convergence in some examples. Section 3 analyses the reasons of the slow convergence of the algorithm in data set with masked high leverage outliers and justifies that this problem does not depend on the particular model used to generate the outliers. Some final comments appear in section 4.

2 GIBBS SAMPLING IN THE SCALE CONTAMINATED MODEL

2.1 Implementation of the Gibbs Sampler

The lack of homogeneity in the sample is frequently modeled with a mixture of distributions. In this paper, we shall focus on identifying outliers in the scale contaminated normal model introduced by Tukey (1960), which has been studied among others by Box and Tiao (1968). In this model, it is assumed that the data may come from a central distribution with high probability, $(1-\alpha)$, and from a contaminated distribution with low probability, α, and that the observations $\boldsymbol{y} = (y_1, \ldots, y_n)'$ are generated by

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \qquad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ are non-random variables; n is the sample size; $\boldsymbol{\beta} \in \mathbf{R}^{p+1}$ is a vector of unknown parameters, and u_i is a random variable with a normal mixture distribution,

$$u_i \sim (1-\alpha)\, N(0, \sigma^2) + \alpha\, N(0, k^2\sigma^2) \qquad i = i, \ldots, n. \tag{2.2}$$

Thus, α is the prior probability that each observation has a $N(\boldsymbol{x}_i'\boldsymbol{\beta}, k^2\sigma^2)$ distribution. We assume that the contamination α and the scale parameter k are known, and also that $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ is a full rank matrix.

The procedure to apply the Gibbs sampling to outlier problems, following Verdinelli and Wasserman (1991), is to introduce a set of dummy variables and compute their posterior probabilities. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$ be a vector of classification variables, defined by

$$\delta_i = \begin{cases} 1 & \text{if } V(y_i) = k^2\sigma^2 \\ 0 & \text{if } V(y_i) = \sigma^2. \end{cases}$$

The marginal posterior probability for the classification variables can be obtained from the expression

$$P(\delta_i = 1 \,|\, \boldsymbol{y}) = \sum_{j_1=0}^{1} \cdots \sum_{j_n=0}^{1} P(\delta_1 = j_1, \ldots, \delta_i = 1, \ldots, \delta_n = j_n \,|\, \boldsymbol{y}). \tag{2.3}$$

The computation of the i marginal probability requires knowing the probabilities of all the possible configurations where $\delta_i = 1$. This means, for example, that for a sample size $n = 40$ we should compute 2^{40} (approximately 10^{12} probabilities) in order to obtain the exact marginal probabilities (2.3). The Gibbs sampling computational advantages seem to be very useful to detect multiple outliers in this problem.

The basic requirement for the Gibbs sampler is to be able to draw samples from the conditional distributions. It is easy to show that the conditional distributions for the parameters in the model (2.1) and (2.2) with non informative priors $p(\boldsymbol{\beta}, \sigma) \propto \sigma^{-1}$, are as follows.

1. For each i, $\delta_i \,|\, \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2$ has a Bernoulli distribution with success probability

$$p_i = \frac{\alpha f_N(u_i/k\sigma)}{\alpha f_N(u_i/k\sigma) + k(1-\alpha) f_N(u_i/\sigma)}, \tag{2.4}$$

   where f_N is the standard normal density function. Conditional to the parameters of the model, the δ's are independent variables.

2. The distribution of the vector $\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\delta}, \sigma^2$ is $N_{p+1}\left(\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\right)$, where

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{y}$$

   and $\boldsymbol{V}$ is a diagonal matrix with elements $v_{ii} = k^{-2}$ if $\delta_i = 1$ and $v_{ii} = 1$ otherwise.

3. The distribution of $\sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{\beta}$ is $Inverted-\chi^2$. Therefore, defining the standarized errors $u_i^* = (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma(1 + \delta_i(k-1))$, it follows that

$$\sum_{i=1}^{n} u_i^{*2} \,|\, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{\beta} \sim \chi_n^2.$$

The Gibbs sampling iterations usually start from an arbitrary vector of initial values $(\sigma^{(0)}, \boldsymbol{\delta}^{(0)}, \boldsymbol{\beta}^{(0)})$. In the first iteration, the samples are generated as follows:

$$\begin{array}{llll} \text{draw} & \sigma_1^{(1)} & \text{from} & \sim \; f(\sigma \,|\, \boldsymbol{y}, \boldsymbol{\delta}^{(0)}, \boldsymbol{\beta}^{(0)}) \\ \text{draw} & \boldsymbol{\delta}^{(1)} & \text{from} & \sim \; f(\boldsymbol{\delta} \,|\, \boldsymbol{y}, \sigma^{(1)}, \boldsymbol{\beta}^{(0)}) \\ \text{draw} & \boldsymbol{\beta}^{(1)} & \text{from} & \sim \; f(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \sigma^{(1)}, \boldsymbol{\delta}^{(1)}). \end{array}$$

Replicating the same scheme s times, we obtain the sequence $(\sigma^{(1)}, \boldsymbol{\delta}^{(1)}, \boldsymbol{\beta}^{(1)}), \ldots, (\sigma^{(s)}, \boldsymbol{\delta}^{(s)}, \boldsymbol{\beta}^{(s)})$. Geman and Geman (1984) have proved that, under regularity conditions, this sequence converges in distribution to $(\sigma, \boldsymbol{\delta}, \boldsymbol{\beta})$. After s iterations and replicating the same scheme r times, it may be possible to make inference for the mean, variance or any other characteristic of the parameter posterior distribution by using the independent and identically distributed samples

$$\begin{array}{c} \sigma_1^{(s)}, \ldots, \sigma_r^{(s)} \\ \boldsymbol{\delta}_1^{(s)}, \ldots, \boldsymbol{\delta}_r^{(s)} \\ \boldsymbol{\beta}_1^{(s)}, \ldots, \boldsymbol{\beta}_r^{(s)}. \end{array}$$

Gelfand and Smith (1990) recommended to use the sample estimate of $p_i = E_{\beta,\sigma^2}\left[P(\delta_i = 1 |\, \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2)\right]$, that is,

$$\hat{p}_{i_{r,s}} = \frac{1}{r} \sum_{j=1}^{r} \frac{\alpha \; f_N((y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_j^{(s)})/k\sigma_j^{(s)})}{\alpha \; f_N((y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_j^{(s)})/k\sigma_j^{(s)}) + k(1-\alpha) \; f_N((y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_j^{(s)})/\sigma_j^{(s)})}. \qquad (2.5)$$

This estimate incorporates the information from an equivalent sample of the other parameters and it is more efficient than the sample mean. This result is proved by Gelfand and Smith (1990) for independent samples, and by Liu *et al.* (1994) in the general case. Alternatively, it is possible to estimate p_i with the last r iterations from an unique sequence as long as we desire. Although running the algorithm only once may save computational time, it has the disadvantage that the samples are identically distributed but not independents. As a result of this, and considering that the space parameter dimension (the sample size plus the parameters in the model) are moderated, in the next examples we always run the Gibbs sampling in parallel sequences and use (2.5) to estimate p_i. In addition, we will see in section 3 that in this problem the Gibbs sampling convergence is very sensible to the initial conditions. By running sequences in parallel we may avoid that the conclusions depend on the selection of only one initial parameter vector. For a most detailed description of the Gibbs sampling performance we refer the reader to Gelfand and Smith (1990) and Casella and George (1992).

2.2 Examples

We analyze the performance of the outlier detection procedure based on the Gibbs sampling in four examples. In the first one it is applied to a much analyzed real data set where the convergence is very fast and the outliers are immediately identified. However, as it is revealed in the next examples, based on real and simulated data, if there are outliers which mask or swamp other observations, the algorithm convergence may not be achieved in a reasonable amount of iterations. In addition, the Gibbs sampling may provide a false idea of the probabilities since the series may be stable around wrong limit values.

The algorithm is always run 1,000 times (in parallel) with different initial values. The last iteration of each performance is used to compute the outlier posterior probability estimates $\hat{p}_{i_{r,s}}$ given by equation (2.5). These probabilities will be represented in the graphs by a bar for each data point. Among the possibilities for selecting the initial values, the designed criterion is to select $\delta_i^{(0)} = 1$ with α probability. Then $\boldsymbol{\beta}^{(0)}$ is the generalized least square estimate (GLS), $\boldsymbol{\beta}^{(0)} = (\boldsymbol{X}'\boldsymbol{V}^{(0)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{(0)}\boldsymbol{y}$, in which $\boldsymbol{V}^{(0)}$ is a diagonal matrix with diagonal elements $1/k^2\sigma^2$ if $\delta_i^{(0)} = 1$, and $1/\sigma^2$ otherwise. It is not necessary to specify the initial value for the variance because it is the first parameter computed in the iterations.

Example 1 The "Stack Loss Data" is a group of real data from a plant for the oxidation of ammonia to nitric acid; 21 diary observations are collected for three explanatory variables and one response variable. This data has been studied with different methods for outlier detection and data 1, 3, 4 and 21 are found to be outliers (see for instance, Daniel and Wood, 1980 or Rousseeuw and Zomeren, 1990). Moreover, some authors add observation

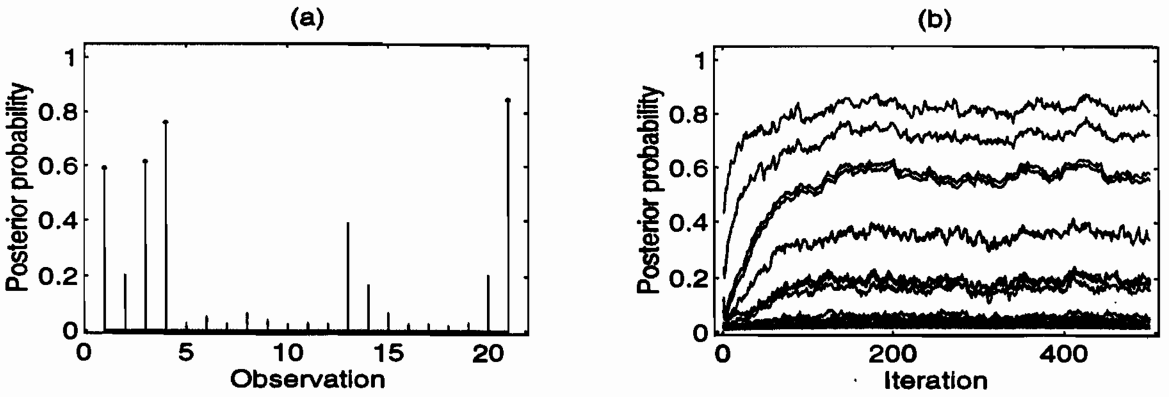

Figure 1: Results of the Gibbs sampler with the Stack Loss data: (a) posterior probabilities for each data point to be outlier after 500 iterations; (b) posterior probabilities as a function of the iteration number.

2 to this list. The data may be found in Daniel and Wood (1980), as well as a description of the experiment.

The outlier posterior probabilities after 500 iterations of the algorithm are represented in Figure 1(a). The results confirm that data 1, 3, 4 and 21 are outliers, with probabilities greater than 0.5. Moreover, the Figure 1(b) shows the series of posterior probabilities for each data as a function of the iteration number. It can be seen that convergence is reached in a few iterations (less than 200).

Example 2 The set of data generated by Hawkins, Bradu and Kass (1984) is a typical example of masking. It includes 75 observations of four variables. Figure 2 shows all the two-dimensional scatter plots that can be obtained by taking pairs of variables. The first fourteen points are high leverage data and of those the first ten are outliers which mask each other and swamp the four non outliers. The outliers will not be easily detected because of the masking and swamping.

After 2,000 iterations of the Gibbs sampling Figure 3(a) shows clearly that the ten outliers are not identified and that it exists a large swamping effect for observations 11 to 14, whose probabilities of being outliers are almost one. The series seems to have

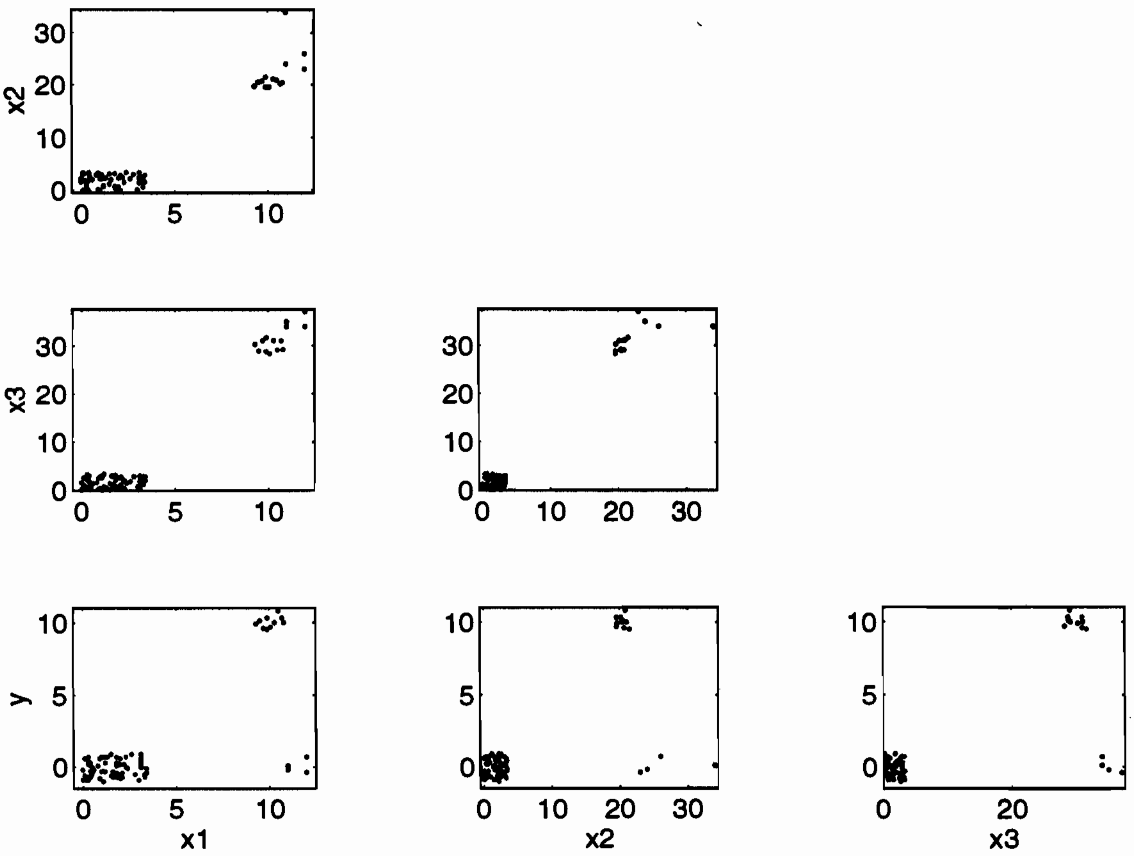

Figure 2: Matrix plot for the Hawkins-Bradu-Kass data.

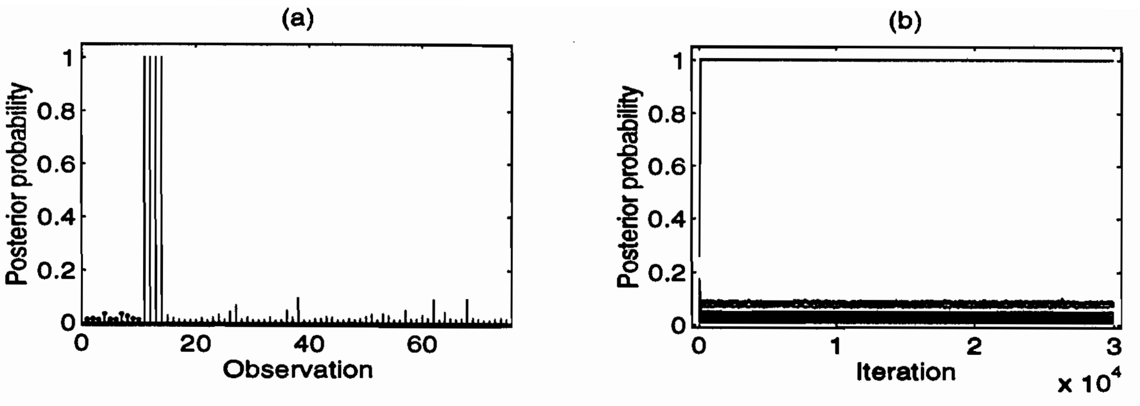

Figure 3: Results of the Gibbs sampler with Hawkins-Bradu-Kass data: (a) posterior probabilities for each data point to be outlier after 2,000 iterations; (b) posterior probabilities as a function of the iteration number.

converged in a few iterations and this wrong result is not modified after 30,000 iterations (see Figure 3(b)).

Example 3 The third set of data is built following Rousseeuw (1984). These are 50 observations with 30 good data points generated from a linear model given by the equation $y_i = 2 + x_i + u_i$, where x_i is a random variable with uniform distribution on (1,4) and the errors are normally distributed with standard deviation 0.2. The 20 outliers are generated from an independently normally distribution with mean vector $\mu = (7, 2)'$ and standard deviations 0.5. The scatter plot of these points is shown in Figure 4, where it can be seen two groups of points. The group on the right correspond to the bad data, observations 1 to 20, that are 40 per cent of the sample.

The final probabilities and the series are shown in Figure 5(a) and Figure 5(b), respectively. After 30,000 iterations, it can be seen that the first 20 observations —the outliers— are not identified when the series seem to converge.

Example 4 The Hertzsprung-Russell diagram of the star cluster CYG OB1 showed in Figure 6 is a real data example. Two variables are observed in 47 stars in the direction of

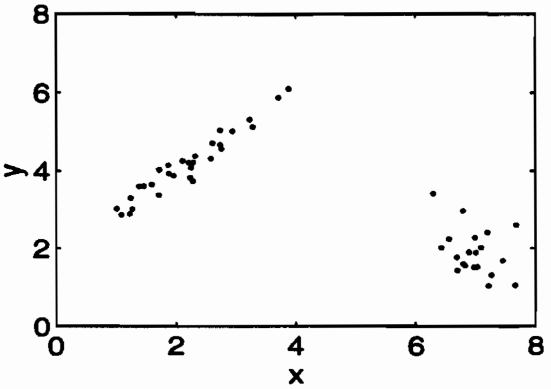

Figure 4: Scatter plot of the Rousseeuw data.

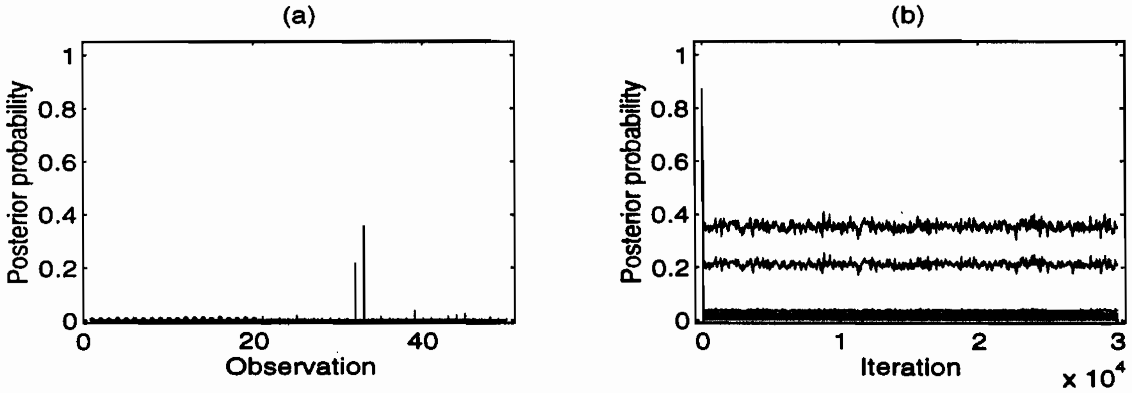

Figure 5: Results of the Gibbs sampling with Rousseeuw data: (a) posterior probabilities for each data point to be outlier after 30,000 iterations; (b) posterior probabilities as a function of the iteration number.

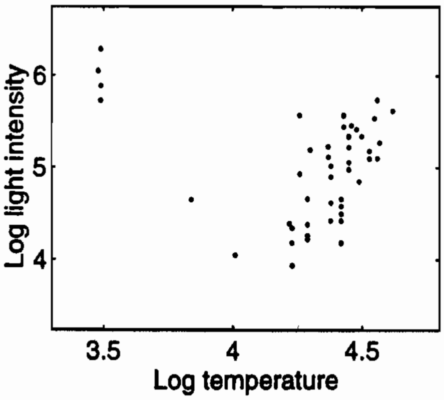

Figure 6: Hertzsprung-Russell diagram of the star cluster CYG OB1.

Cygnus. The independent variable (x) is the logarithm of the effective temperature at the surface of the star and the dependent variable (y) is the logarithm of the light intensity. The values are provided by Rousseeuw and Leroy (1987). The scatter plot shows that exist four outliers (observations 11, 20, 30 and 34) which correspond with giant stars.

This example shows that the convergence problem observed in the previous examples may also appear in real data sets. It can be seen in Figure 7(a) and Figure 7(b) that after 10,000 iterations the outliers are not identify and the series seem to converge.

3 ANALYSIS OF THE GIBBS SAMPLING CONVERGENCE

The examples in the previous section have shown that the direct application of the Gibbs sampling will be a bad procedure for outliers detection in certain data sets, because the posterior probability series may seem to converge around false values.

One reason for this is the masking problem. If outliers mask or swamp each other, their δ variables are high correlated and, also, the parameter space dimension (the sample size plus the parameters in the model) rises with the sample size. Smith and Roberts (1993) indicated that high dimensional parameter space and high correlation will slow down the convergence, but the problem is more serious that the one indicated by these authors.

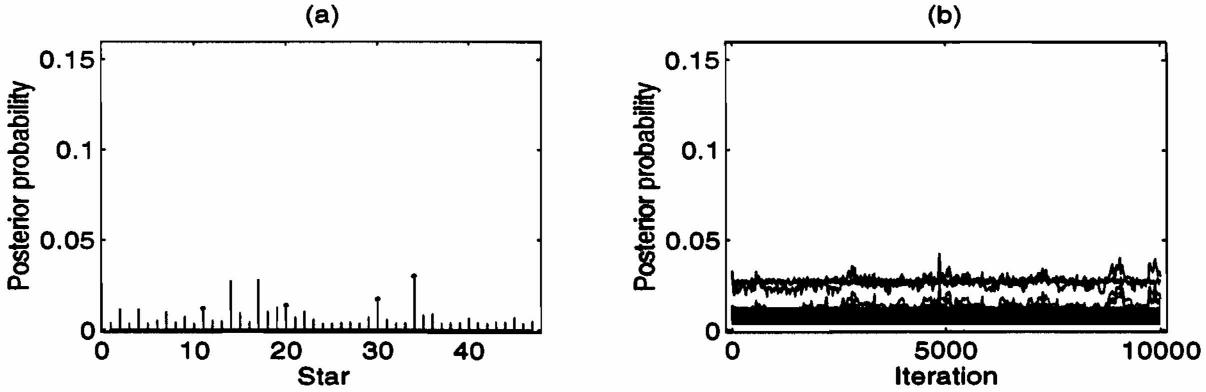

Figure 7: Results of the Gibbs sampling with data of the Hertzsprung-Russell diagram of the star cluster CYG OB1: (a) posterior probabilities for each data point to be outlier after 10,000 iterations; (b) posterior probabilities as a function of the iteration number.

For instance, the data in Figure 8 is a sample of a two normal mixture (contamination is thirty percent of the data) in which these two conditions will appear. The probabilities in Figure 9(a) and the series in Figure 9(b) show that the convergence is slow, as expected, but it is eventually achieved. This is not the case in the regression examples in section 2.2. The principal difference among these two situations is the role that leverage plays in the regression model. If the initial assignation of the classification variables includes as good data points many of the high leverage outliers which cause masking and/or swamping, the regression coefficients will be biased, the residuals at these points will be very small, and the probability of these points to be classified as outliers will be low in the next iterations.

Let $\boldsymbol{\delta}^{(0)}$ be the initial configuration to start the algorithm and let $\boldsymbol{\delta}^{(0)}$ and let $\boldsymbol{\beta}^{(0)}$ be the generalized least square estimate using $\boldsymbol{\delta}^{(0)}$. In the first iteration, $\delta_i^{(1)} = 1$ with probability $p_i^{(1)}$ given by (2.4), in which $\boldsymbol{\beta}$ is substituted by $\boldsymbol{\beta}^{(0)}$ and σ by the standard deviation drawn in the first iteration. The probability $p_i^{(1)}$ can be expressed as

$$p_i^{(1)} = \left(1 + k\,\alpha^{-1}(1-\alpha)\,\exp\left(-\frac{1}{2\phi^{-1}\sigma^{2(1)}}u_i^{(0)2}\right)\right)^{-1}, \qquad (3.1)$$

where $u_i^{(0)} = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}^{(0)}$ and $\phi = 1 - k^{-2}$. For large k, the probability (3.1) only depends

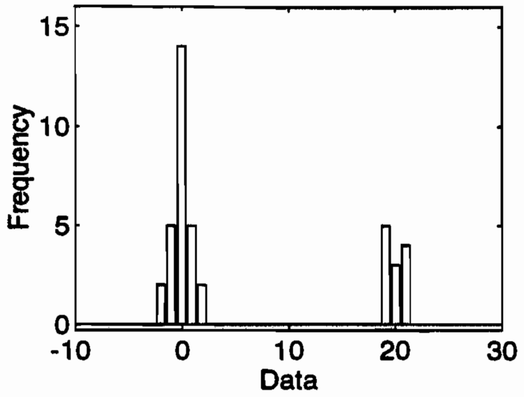

Figure 8: (a) Frequency histogram of $n = 40$ data generated from a normal mixture distribution.

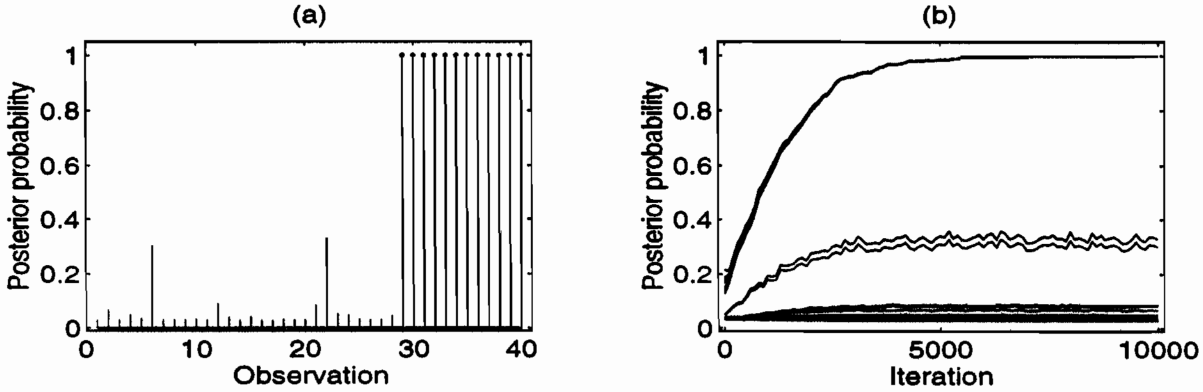

Figure 9: Results of the Gibbs sampling with data generated from a mixture normal distribution; (a) posterior probabilities for each data point to be outlier after 10,000 iterations; (b) posterior probabilities as a function of the iteration number.

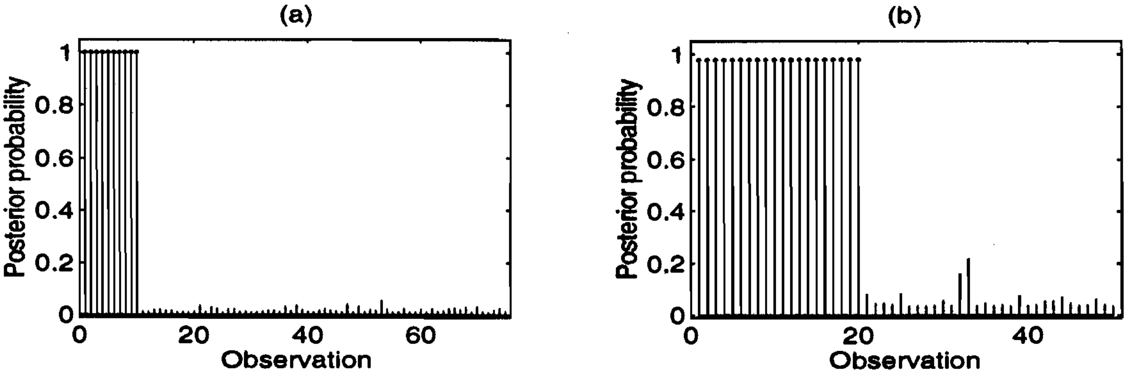

Figure 10: Posterior probabilities after 200 iterations when the outliers are initially assigned to the contaminated distribution: (a) Hawkins-Bradu-Kass data; (b) Rousseeuw data.

on the residual $u_i^{(0)}$ ($\sigma^{(1)}$ is the same for all of them) and it will be close to one when $u_i^{(0)}$ is large, and close to zero when $u_i^{(0)}$ is small.

Let $\boldsymbol{S}_0 = (\boldsymbol{X}_0, \boldsymbol{y}_0)$ be the observations that are classified as good in the initial conditions. For large k, $\boldsymbol{V}^{(0)}$ is approximately the identity matrix and, therefore, $u_i^{(0)}$ will be the least square residual using the subsample $(\boldsymbol{X}_0, \boldsymbol{y}_0)$. If this subsample contains several high leverage outliers, the coefficient $\boldsymbol{\beta}^{(0)}$ will be biased and the least square residuals at these points will be small. Therefore, they will have a very low probability of being selected as outliers in the next iteration. The only chance of detecting these outliers will be when all of them are classified as outliers in the drawing from the conditional distribution (3.1). For instance, if we have 10 outliers and $p_i^{(1)} = 0.01$, this probability is 10^{-20}.

The solution to this problem begins with the correct initial assignation of the group of masked outliers. For the examples 2 and 3 analysed in section 2.2, the graphs in Figure 10 show the probabilities when, at least, the outliers are initially assigned to the contaminated distribution. As it can be seen, convergence is reached very quick.

One may wonder if the lack of convergence shown in the examples is due to the particular model used. For instance, instead of the scale contaminated model (2.1) and (2.2) we may have assumed the mean-shift model utilized by Guttman (1973) and Guttman, Dutter and Freeman (1978) or, even, assume no particular model for the generation of the bad data, as advocated by Geisser (1991) and Pettit and Smith (1985). However, as shown by Peña and Guttman (1993) for large k, as assumed in this paper, the probabilities computed by the Tukey (1960) model, the mean-shift model and the predictive approach, in which no model for the generation of the outliers is used, are essentially the same. The reason is that for large k, model (2.1) and (2.2) allows any departure from the central model, which is equivalent to allowing any mean-shift or any source of heterogeneity (see also Guttman and Peña, 1993).

We have also considered a most general non-parametric hierarchical model. In this model, the observations are generated by the equation (2.1) but now the error distributions are

$$u_i \sim (1-\alpha)\, N(0,\sigma^2) + \alpha\, N(h_i, \sigma^2\tau_i^2) \qquad i = 1,\ldots,n. \tag{3.2}$$

As different level and scale parameters for the contaminated distribution have to be estimated using only one observation, the model is unidentified, except when some observations share a common parameter. For this to happen, the distribution of the pairs $\theta_i = (h_i, \tau_i^2)$ should be discreet. Therefore, to complete the prior structure we consider the following distributions:

$$\begin{aligned} \theta_i &\sim G \\ G &\sim \textit{Dirichlet Process}\ (\mu, G_0) \\ G_0 &\sim N(m,b) \times \textit{Inv} - \textit{Gamma}\ (u/2, v/2) \\ \mu &\sim \textit{Gamma}\ (a_0, b_0), \end{aligned}$$

where G is an unknown bivariate distribution, μ is the total mass and G_0 is the prior expectation of the Dirichlet Process (Ferguson, 1973).

Escobar (1994) proposed the use of Gibbs sampling in problems which involve Dirichlet process priors and showed that

$$\theta_i \mid \boldsymbol{y}, \theta_{(i)} \sim \pi_{n+1} G_i + \sum_{j \neq i} \pi_j \ I_{(\theta_i = \theta_j)}, \tag{3.3}$$

where $\theta_{(i)} = (\theta_1, \ldots, \theta_{n-1}, \theta_{n+1}, \ldots, \theta_n)$, $\pi_{n+1} + \sum_{j \neq i} \pi_j = 1$, and I_A is the unit point mass at A. The equation (3.3) means that in the Gibbs sampling iterations the parameter

θ_i is one of the values in $\theta_{(i)}$ with probability $\pi_j \propto f(y_i \mid \theta_j)$, and with probability $\pi_{n+1} \propto \int f(y_i \mid \theta)\, dG_0(\theta)$ is drawn from G_i, that is the posterior distribution of θ_i given the data y_i and the prior distribution G_0. Nevertheless, we use the modified scheme of the Gibbs sampling introduced by MacEachern (1994) and implemented by Müller, Erkanli and West (1992) in the nonparametric estimation of the regression function. The parameter vector is augmented with n group indicators $\boldsymbol{s} = (s_1, \ldots, s_n)$ which hold that $s_i = s_{i'} = j$ if and only if $\theta_i = \theta_{i'} = \theta_j^\star$, where $\boldsymbol{\theta}^\star = (\theta_1^\star, \ldots, \theta_k^\star)'$ is the vector of the $k \leq n$ distint values in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$. The posterior distributions for $\boldsymbol{\delta}$, $\boldsymbol{\beta}$ and σ^2 have the same structure than in model (2.1) and (2.2) and are given in the appendix, as well as the conditional distributions of $\boldsymbol{s}$, $\boldsymbol{\theta}^\star$ and μ.

We have applied this model to the examples in section (2.2), finding the same results that are shown there in all the four cases.

4 CONCLUDING REMARKS

The Gibbs sampling can be used for outlier detection as Verdinelli and Wasserman (1991) showed in the estimation of the mean for a normal model. When outliers are isolated, Gibbs sampling avoids the 2^n necessary computation to obtain the marginal posterior probabilities in the scale contaminated regression model. However, when the set of data has many outliers that mask each other, Gibbs sampling will fail and posterior distributions are badly estimated. An erroneous initial classification of the observations will conduct the algorithm to a wrong solution along thousands of iterations. The examples have shown that in regression high leverage may avoid convergence completely.

APPENDIX: CONDITIONAL DISTRIBUTIONS FOR THE NONPARAMETRIC MODEL (2.1) AND (3.2)

The conditional distributions for the parameters in the model (2.1) and (3.2) are as follows:

1. For each i, $\delta_i \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{s}, \theta^\star$ has a Bernoulli distribution with success probability
$$p_i = \frac{\alpha f_N((u_i - h^\star_{s_i})/\sigma\tau^\star_{s_i})}{\alpha f_N((u_i - h^\star_{s_i})/\sigma\tau^\star_{s_i}) + (1-\alpha)\tau^\star_{s_i} f_N(u_i/\sigma)}.$$
2. The distribution of the vector $\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{s}, \boldsymbol{\theta}^\star$ is $N_{p+1}(\hat{\boldsymbol{\beta}}_s, \sigma^2(\boldsymbol{X}'\boldsymbol{V}_s\boldsymbol{X})^{-1}))$, where $\hat{\boldsymbol{\beta}}_s = (\boldsymbol{X}'\boldsymbol{V}_s\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}_s(\boldsymbol{y} - \boldsymbol{H}_s)$, $\boldsymbol{H}_s = (\delta_1 h^\star_{s_1}, \ldots, \delta_n h^\star_{s_n})'$ and $\boldsymbol{V}_s$ is a diagonal matrix with elements $(1 + \delta_i(\tau^{2\star}_{s_i} - 1))^{-1}$.
3. The distribution of $\sigma^2 \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{s}, \boldsymbol{\theta}^\star$ is $Inverted - Gamma\ (n/2, \sigma^2_s/2)$, where $\sigma^2_s = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{H}_s)'\boldsymbol{V}_s(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{H}_s)$.
4. Let $\boldsymbol{s}_{(i)}$ be the vector $\boldsymbol{s}$ when s_i is eliminated and let n_{ij} be the number of group indicators in $\boldsymbol{s}_{(i)}$ equal to j. Then the number of different indicators is
$$k_{(i)} = \begin{cases} k-1 & \text{if } s_i \neq s_j \text{ and } j \neq i \\ k & \text{otherwise.} \end{cases}$$
In order to compute $\pi_{i,j} = P(s_i = j \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{s}_{(i)}, \boldsymbol{\theta}^\star, \mu)$ we consider two cases:
   - (i) When $\delta_i = 1$, the probability $\pi_{i,j}$ is given by
$$\pi_{i,j} = \begin{cases} C\ n_{ij}\ \tau^\star_j\ f_N((u_i - h^\star_j)/\sigma\tau^\star_j) & \text{for } j = 1, \ldots, k_{(i)} \\ C\ \mu\ \tau^\star_{s_i}\ f_N((u_i - h^\star_{s_i})/\sigma\tau^\star_{s_i}) & \text{for } j = k_{(i)} + 1, \end{cases}$$
where $C = (\pi_{i,k_{(i)}+1} + \sum_{j \neq i} \pi_{i,j})^{-1}$. Note that $\pi_{i,k_{(i)}+1}$ is proportional to $\int f(y_i \mid \theta) dG_0(\theta)$ and it is approximated by the density of a $N(\boldsymbol{x}'_i\boldsymbol{\beta} - h^\star_{s_i}, \sigma^2\tau^{2\star}_{s_i})$.
   - (ii) When $\delta_i = 0$, the probability $\pi_{i,j}$ is given by
$$\pi_{i,j} = \begin{cases} n_{ij}/(\mu + n - 1) & \text{for } j = 1, \ldots, k_{(i)} \\ \mu/(\mu + n - 1) & \text{for } j = k_{(i)} + 1. \end{cases}$$
5. For $j = 1, \ldots, k$, we define the sets $I^\star_j = \{i \mid \delta_i = 1 \text{ and } \ s_i = j\}$ and call $n^\star_j$ to the size of $I^\star_j$. Then the conditional distributions of $h^\star_j$ and $\tau^\star_j$ are:
$$\begin{aligned} h^\star_j \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{s}, \tau^{2\star}_j &\sim N(m_j, b_j) \\ \tau^{2\star}_j \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{s}, h^\star_j &\sim Inverted - Gamma\ \left(\frac{n^\star_j + u}{2}, \frac{v + v_j}{2}\right), \end{aligned}$$
where $b_j = (b^{-2} + \tau_j^{-2\star}\sigma^{-2}n^\star_j)^{-1}$, $m_j = b_j\left(b^{-2}m + \tau_j^{-2\star}\sigma^{-2}\sum_{i \in I^\star_j} u_i\right)$ and $v_j = \sigma^{-2}\sum_{i \in I^\star_j}(u_i - h^\star_j)^2$.

6. The conditional distribution of μ is computed by augmenting the parameter vector with an artificial variable η (see Escobar and West, 1995). The conditional distributions are given by

$$\eta \mid \boldsymbol{y}, \mu \sim Beta(\mu + 1, n)$$
$$\mu \mid \boldsymbol{y}, \boldsymbol{s}, \eta \sim \pi \; Gamma(a_1, b_1) + (1 - \pi) \; Gamma(a_1 - 1, b_1),$$

where $\pi = (a_1 - 1)/(a_1 - 1 + nb_1)$, $a_1 = a_0 + k$ and $b_1 = b_0 - log(\eta)$.

ACKNOWLEDGEMENTS

The first author thanks Mike West and Christian Robert for useful discussion on this subject when she was visiting the Institute of Statistics and Decision Sciences, Duke University, and the Center for Research in Economics and Statistics, INSEE. Both authors acknowledge support for this research provided by DGICYT (Spain), under grant No. PB93–0230.

REFERENCES

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

Casella, G. and George, E.I. (1992). "Explaining the Gibbs sampler". *American Statistician,* 46, 167–174.

Escobar, M.D. (1994). "Estimating normal means with a Dirichlet process prior". *Journal of the American Statistical Association,* 89, 268–277.

Escobar, M.D. and West, M. (1995). "Bayesian density estimation and inference using mixtures". *Journal of the American Statistical Association,* (to appear).

Ferguson, T.S. (1973). "A Bayesian analysis of some nonparametric problems". *Annals of Statistics,* 1, 209–230.

Geisser, S. (1991). "Diagnostics, divergences and perturbation analysis". *The IMA Volumes in Mathematics and its Applications,* 33(1), 89–100, ed. W. Stahel and S. Weisberg, Springer Verlag.

Gelfand, A.E. and Smith, A.F.M. (1990). "Sampling-based approaches to calculating marginal densities". *Journal of the American Statistical Association,* 85, 398–409.

Gelman, A. and Rubin, D.B. (1992). "A single series from the Gibbs sampler provides a false sense of security". *Bayesian Statistics 4,* 625–631, ed. J. Bernardo *et al.*, Oxford University Press.

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

Guttman, I. (1973). "Care and handling of univariate or multivariate outliers in detecting spuriousity — A Bayesian approach. *Technometrics,* 15, 723–738.

Guttman, I., Dutter, R. and Freeman, P.R. (1978). "Care and handing of univariate outliers in the general linear model to detect spuriousity — A Bayesian approach". *Technometrics,* 20, 187–193.

Guttman, I. and Peña, D. (1993). "A Bayesian look at diagnostic in the univariate linear model". *Statistica Sinica,* 3, 367–390.

Hawkins, D.M., Bradu, D. and Kass, G.V. (1984). "Location of several outliers in multiple regression data using elemental sets". *Technometrics,* 26, 197–208.

Hills, S.E. and Smith, A.F.M. (1991). "Parametrization issues in Bayesian inference". *Bayesian Statistics 4,* 227–246, ed. J. Bernardo *et al.*, Oxford University Press.

Liu, J., Wong, W.H. and Kong, A. (1994). "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes". *Biometrika,* 81, 27–40.

MacEachern, S.N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior". *Communications in Statistics, Simulation and Computing,* 23, 727–741.

Matthews, P. (1993). "A slowly mixing Markov chain with implications for Gibbs sampling". *Statistics and Probability Letters,* 17, 231–236.

Mengersen, K.L. and Robert, C.P. (1994) "Testing for mixtures: a Bayesian entropic approach". *Bayesian Statistics 5* (in press).

Müller, P. Erkanli, A. and West, M. (1992). "Bayesian curve fitting using multivariate normal mixtures". ISDS Discussion Paper 92–A09, Duke University.

Peña, D. and Guttman, I. (1993). "Comparing probabilistic methods for outlier detection in linear models". *Biometrika,* 80, 603–610.

Pettit, L.I. and Smith, A.F.M. (1985). "Outliers and influential observations in linear models". *Bayesian Statistics 2,* 473–494, ed. J. Bernardo *et al.*, Oxford University Press.

Polson, N.G. (1994). "Convergence of Markov Chain Monte Carlo algorithms". *Bayesian Statistics 5* (in press).

Rousseeuw, P.J. (1984). "Least median of squares regression". *Journal of the American Statistical Association,* 79, 871–880.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier detection.* New York: John Wiley.

Smith, A.F.M. and Roberts, G.O. (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods" (with discussion). *Journal of the Royal Statistical Society, B,* 55, 3–24.

Tukey, J.W. (1960). "A survey of sampling from contaminated distributions". *Contributions to Probability and Statistics: Volume Dedicated to Harold Hotelling,* Stanford: University Press.

Verdinelli, I. and Wasserman, L. (1991), "Bayesian analysis of outlier problems using the Gibbs sampler". *Statistics and Computing,* 1, 105–117.