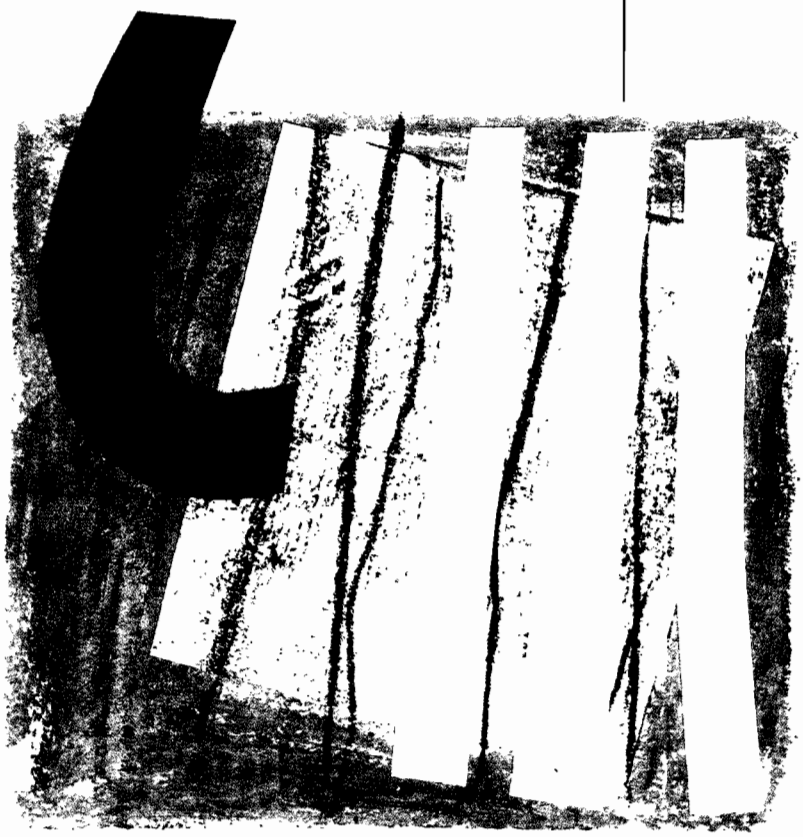
BAYESIAN UNMASKING IN LINEAR MODELS

Ana Justel and Daniel Peña

96-47



WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 96-47
Statistics and Econometrics Series 18
September 1996

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

BAYESIAN UNMASKING IN LINEAR MODELS

Ana Justel and Daniel Peña*

Abstract

We propose a Bayesian procedure for multiple outlier detection in linear models avoiding the masking problem. Our proposal is illustrated with several examples in which our procedure outperforms other recent methods for multiple outlier detection. The posterior probabilities of each data point being an outlier are estimated by using a new adaptive Gibbs sampling method, which modifies the initial conditions of the Gibbs sampler by using the eigenstructure of the covariance matrix of the indicator variables. This procedure also overcomes the false convergence of the Gibbs sampling in problems with strong masking.

Key Words

Gibbs sampler; linear regression; multiple outliers; sequential learning.

*Centre for Operations Research and Econometrics, Universitè Catholique de Louvain; Department of Statistics and Econometrics, Universidad Carlos III de Madrid.

BAYESIAN UNMASKING IN LINEAR MODELS

Ana Justel* and Daniel Peña**

*Center for Operations Research and Econometrics,
Universitè Catholique de Louvain
**Department of Statistics and Econometrics,
Universidad Carlos III de Madrid

Abstract

We propose a Bayesian procedure for multiple outlier detection in linear models avoiding the masking problem. Our proposal is illustrated with several examples in which our procedure outperforms other recent methods for multiple outlier detection. The posterior probabilities of each data point being an outlier are estimated by using a new adaptive Gibbs sampling method, which modifies the initial conditions of the Gibbs sampler by using the eigenstructure of the covariance matrix of the indicator variables. This procedure also overcomes the false convergence of the Gibbs sampling in problems with strong masking.

Key words: Gibbs sampler. Linear regression. Multiple outliers. Sequential learning.

1 INTRODUCTION

Diagnostic methods for identifying a single outlier or influential observation in a linear model are well established in the statistical literature either from the Classical or Bayesian point of view. See Cook and Weisberg (1982), Pettit and Smith (1985) and Peña and Guttman (1993). However, the identification of multiple outliers in linear models is a difficult problem because the masking effect. Some recent proposals to solve the problem from the Classical point of view are Hadi and Simonoff (1993) and Peña and Yohai (1995). Rousseeuw and Zomeren (1990) and Atkinson (1994) have proposed the use of robust estimation to identify multiple outliers.

This paper presents a new procedure based on the Bayesian approach to identify multiple outliers in linear models. The proposed method seems to work better than other procedures recently presented in the literature. The posterior probabilities of each observation being an outlier are computed by an adaptive Gibbs sampling procedure that overcomes problems of convergence due to the masking effect. The result is a two stage method which seems to work very well in problems with multiple outliers and strong masking. The first stage uses a few iterations of the Gibbs sampling and the information available when the series of outlier probabilities are stable to determinate the initial conditions in the second stage.

The paper is organized as follows. In section 2 the model and a brief review of the literature on outliers in Bayesian linear models is presented. Section 3 develops the new adaptive procedure. Section 4 applies it to some examples with real and simulated data, showing its good performance in samples with masking and swamping problems. The procedure is compared to the outlier detection methods by Hadi and Simonoff (1993) and Peña and Yohai (1995) finding that it works where these other methods may fail. Some final comments appear in section 5.

2 OUTLIERS IN THE BAYESIAN LINEAR MODEL

Let us consider the Bayesian regression model where 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 n is the sample size, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ is a $n \times p$ matrix of non random variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{u} = (u_1, \ldots, u_n)'$ is a vector of non observable perturbations with distribution $N(0, \sigma^2)$. We assume independent and non informative prior distributions for the location and scale parameters, $P(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$. Bayesian methods for outlier detection can be classified into two groups: (1) diagnostic methods which propose a null model for the data generation excluding that outliers may be generated; and (2) robust methods which propose a model for the generation of all the data set, including the possible outliers.

The diagnostic methods analyze if one observation is compatible with the rest of the sample by studying the predictive distribution $p(y_i \mid \boldsymbol{y}_{(i)})$, where $\boldsymbol{y}_{(i)}$ is the sample excluding the data y_i. This measure is called the conditional predictive ordinate method (Geisser, 1980 and Pettit and Smith, 1985) and Pettit (1990) proves that it is related to the studentized residual test. In this case the predictive ordinate is given by

$$p(y_i \mid \boldsymbol{y}_{(i)}) = c\, s_{(i)}^{-1}(1-h_i)^{1/2}\left(1 + \frac{t_i^2}{n-p-1}\right)^{-\frac{n-p}{2}} \tag{2.2}$$

where t_i is the studentized residual, $s_{(i)}^2$ is the unbiased estimate of σ^2 when the data y_i is eliminated, and h_i is the ith element in the principal diagonal of the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Therefore, data with large studentized residual will have a small predictive ordinate and will be consider outliers.

The robust methods suppose heavy tail distributions for the errors or mixtures of distributions (e.g. Box and Tiao, 1973 or West, 1984). The more frequently analyzed model is the normal scale contamination model, where the error distribution is

$$u_i \sim (1-\alpha)\; N(0, \sigma^2) + \alpha\; N(0, k^2\sigma^2) \qquad i = 1, \ldots, n. \tag{2.3}$$

Assuming that k and α are known, the posterior probability that there are n_{I} outliers in a set indexed by $I = \{i_1, \ldots, i_{n_{\mathrm{I}}}\}$ is given by

$$p_{\mathrm{I}} \propto \left(\frac{\alpha}{1-\alpha}\right)^{n_{\mathrm{I}}} k^{-n_{\mathrm{I}}} \left(\frac{\mid \boldsymbol{X}'\boldsymbol{X} \mid}{\mid \boldsymbol{X}'\boldsymbol{X} - \phi\boldsymbol{X}_{\mathrm{I}}'\boldsymbol{X}_{\mathrm{I}} \mid}\right)^{\frac{1}{2}} \left(\frac{s^2}{s_{(\mathrm{I})}^2}\right)^{\frac{n-p}{2}} \tag{2.4}$$

where $\phi = 1 - k^{-2}$, $\boldsymbol{X}_{\mathrm{I}}$ is the $n_{\mathrm{I}} \times p$ submatrix of $\boldsymbol{X}$ with the rows indexed by I, s^2 is the usual unbiased residual variance estimate and $s_{(\mathrm{I})}^2$ is computed by considering the

n_I points in I generated from the alternative distribution. The model (2.1) and (2.3), introduced by Tukey (1960), has been studied among others by Box and Tiao (1968), Freeman (1980), Pettit (1992) and Peña and Tiao (1992). The mixture distribution (2.3) indicates that it exists a probability α of each data point being spuriously generated from an alternative distribution. Data points generated from the alternative distribution will be consider outliers. The advantage of this model with respect to the heavy tail ones is that it not only produces an efficient robust parameter estimation but also it can provide an outlier identification procedure. When k is large it can be shown (Peña and Guttman, 1993) that the behavior of this model for outlier identification is similar to the mean-shift model by Guttman (1973) and to the predictive ordinate method (2.2).

The formulas (2.2) and (2.4) can be easily used to check for a single outlier in the sample. However, when the number and the position of outliers are unknown, that is the usual case with real data, two detection procedures has been proposed: (1) using the deleting one observation procedure to detect outliers one by one; and (2) considering multiple detection for identifying groups of outliers.

The deleting one observation procedures with multiple outliers can be subject to masking. Masking occurs when one outlier observation is not detected because of the presence of others outliers. Also, one good point can be wrongly identified as outlier due to the effect of the outliers, and this is called the swamping problem. The multiple detection procedures using (2.4) may avoid masking, but they involve the extensive computations of the 2^n posterior probabilities which correspond to all the possible configurations for the generation of the data. Peña and Tiao (1992) propose a method based on stratified sampling to reduce the computations in the context of building the Bayesian robustness curves BROC and SEBROC. Verdinelli and Wasserman (1991) apply the Gibbs sampling algorithm (Geman and Geman, 1984 and Gelfand and Smith, 1990) to the detection of univariate outliers in a normal random sample and show that this algorithm overcomes the heavy computations needed in this type of problems. Justel and Peña (1996) extend the procedure to the outlier detection in linear regression and show that, when the outliers are isolated, Gibbs sampling works well and avoids the 2^n necessary computation to obtain the marginal posterior proba-

bilities. However, in strong masking cases the algorithm fails and multiple outliers are not always detected when the convergence seems to be reached. The fault is attributed to the problems of high contamination and the presence of influence outliers in the sample.

In this paper we generalize the normal scale contamination model (2.1) and (2.3) by assuming for the contamination parameter α a prior distribution $Beta(\gamma_1, \gamma_2)$ with expectation $\alpha_0 = E(\alpha) = \gamma_1/(\gamma_1 + \gamma_2)$. The application of the Gibbs sampling is carried out by augmenting the parameter vector with a set of classification variables $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$, that are binary variables defined as $\delta_i = 1$ if y_i is generated by the alternative distribution $N(\boldsymbol{x}_i'\boldsymbol{\beta}, k^2\sigma^2)$, and $\delta_i = 0$ otherwise. The pair $(y_i, \boldsymbol{x}_i')$ will be called an outlier when the marginal posterior probability p_i that its classification variable is equal to one is greater than 0.5. Thus, α is the prior probability that any observation is an outlier. Then the full conditional distributions are: (1) the conditional distribution of $\boldsymbol{\beta}$ is $N_p\left(\tilde{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\right)$, where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$ and $\boldsymbol{V}$ is a diagonal matrix with elements $v_{ii} = k^2$ if $\delta_i = 1$ and $v_{ii} = 1$ otherwise; (2) the conditional distribution of σ^2 is $Inverted - Gamma\ \left(n/2, \sum u_i^{*2}/2\right)$, where $u_i^* = (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})/(1+\delta_i(k-1))$; (3) the conditional distribution of δ_i is $Bernoulli$ with success probability

$$P(\delta_i = 1 \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \alpha) = \frac{\exp\left(-u_i^2/2k^2\sigma^2\right)\alpha}{\exp\left(-u_i^2/2k^2\sigma^2\right)\alpha + \exp\left(-u_i^2/2\sigma^2\right)(1-\alpha)k}; \tag{2.5}$$

and (4) the conditional distribution of α only depends on the vector $\boldsymbol{\delta}$ and is $Beta\ (\gamma_1 + n\bar{\delta}, \gamma_2 + n(1-\bar{\delta}))$, where $\bar{\delta} = \sum \delta_i/n$. Note that the conditional expectation is a linear combination of the prior expectation and the sample mean

$$E(\alpha \mid \boldsymbol{\delta}) = \frac{\gamma_1 + \gamma_2}{\gamma_1 + \gamma_2 + n}\alpha_0 + \frac{n}{\gamma_1 + \gamma_2 + n}\bar{\delta}.$$

When the Gibbs sampler is run R times, inference for the mean, variance or any other characteristic of the posterior distributions is made by using the independent and identically distributed samples obtained from the last iteration of each performance. In particular, the estimates of the marginal outlier posterior probabilities are

$$\hat{p}_{i_{\mathrm{R}}}^{(\mathrm{S})} = \frac{1}{R}\sum_{r=1}^{\mathrm{R}} \delta_{i_r}^{(\mathrm{S})}, \tag{2.6}$$

and the series of posterior probability estimates (2.6) for each data point, as a function of the iteration number, will be used for monitoring convergence.

Justel and Peña (1996) showed in several examples that Gibbs sampling will fail for outlier detection in data sets with masking problems. A key factor to explain the lack of convergence in these cases seems to be the effect of the leverage in the estimation of linear regression models. When high leverage outliers which cause masking are classified as good data in the initial vector $\boldsymbol{\delta}^{(0)}$, the probability that these points are identified as outliers depends on the residuals $u_i^{(0)} = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}^{(0)}$ and will be low in the next iterations. This fact can be easily seen in the extreme case in which the sample includes a group indexed by I of n_{I} identical outliers. Let $\boldsymbol{S}_0 = (\boldsymbol{y}_0, \boldsymbol{X}_0)$ be the set of observations classified as good in the initial conditions and let us consider the case in which $\boldsymbol{S}_0$ includes the group of outliers. The probability defined by (2.5) can be expressed in the first iteration as

$$P(\delta_i^{(1)} = 1 \mid \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \alpha) = \left(1 + \left(\frac{1-\alpha^{(1)}}{\alpha^{(1)}}\right) F_{10}^{(1)}(i)\right)^{-1}, \tag{2.7}$$

where $F_{10}^{(1)}$ is the Bayes factor given by

$$F_{10}^{(1)}(i) = k \cdot \exp\left(-\frac{1}{2\phi^{-1}\sigma^{2(1)}} u_i^{(0)2}\right)$$

and $\phi = 1 - k^{-2}$. Peña and Yohai (1995) proved that $u_i^{(0)}$ can be expressed as

$$u_i^{(0)} = \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_{(\mathrm{I})}^{(0)}}{1 + n_{\mathrm{I}}h} \qquad \text{for } i \in I, \tag{2.8}$$

where $h = \boldsymbol{x}_i'(\boldsymbol{X}_{0(\mathrm{I})}'\boldsymbol{X}_{0(\mathrm{I})})^{-1}\boldsymbol{x}_i$ are the common out-of-sample leverage for $i \in I$ and, from now on, the subscript (I) means that the data indexed by I are deleted. For large k, $\boldsymbol{\beta}_{(\mathrm{I})}^{(0)}$ may be approximate by the least square estimate when the observations indexed by I are deleted from $\boldsymbol{S}_0$, that is $\hat{\boldsymbol{\beta}}_{0(\mathrm{I})} \approx (\boldsymbol{X}_{0(\mathrm{I})}'\boldsymbol{X}_{0(\mathrm{I})})^{-1}\boldsymbol{X}_{0(\mathrm{I})}'\boldsymbol{y}_{0(\mathrm{I})}$. It is immediate from equation (2.8) that the residual $u_i^{(0)}$ will be small if h is large (note that h is not bounded) and this effect increases with the number of outliers n_{I}. Therefore, for high leverage outliers the residual $u_i^{(0)}$ will be close to zero and the probability (2.7) will also be close to zero. On the other hand, if the set $\boldsymbol{S}_0$ does not contain outliers, the out-of-sample residuals $u_i^{(0)}$ will be large for $i \in I$ and the probability (2.7) will be

close to one. Therefore, we conclude that the set of outliers will be detected in the next iteration only when all of them are classified as such in the drawing from the conditional distribution (2.7).

3 PROCEDURE TO AVOID MASKING

We have seen in section 2 that when the sample contains a set of masked outliers and the initial set S_0 includes some of these points, the Gibbs sampler is expected to fail. As a result of this analysis it is reasonable to assume the following *initial condition dependence property:*

i) if S_0 includes no outliers, the existing outliers are always identified, and the good data are not misspecified;

ii) if S_0 includes several influential outliers, the probability of identifying all the outliers in the sample is small and will be very close to zero if the number of misspecified outliers is large.

Therefore a clear objective is to start the procedure with a set S_0 that is outlier free. This idea is similar to the one used in robust estimation procedures based on resampling (Rousseeuw, 1984, and Hawkins, Bradu and Kass, 1984). Before starting the algorithm the only information that can be used to build S_0 is that, by definition, outliers will be some small fraction of the data. However, when the Gibbs sampler is run and the outlier probability series stabilize we have information about the dependency among the classification variables. Based on this idea we propose an adaptive-learning method in which the initial conditions of the Gibbs sampler are changed according to a two-stage procedure. In the first stage, the Gibbs sampling is initialized by (i) using a small set of initial values as good observations and (ii) applying diagnostic test to these initial values to eliminate single outliers. Then the algorithm is run for a few iterations until the outlier probability series are stable. The dependency among the classification variables computed from the run is taken into account in order to divide the sample into two groups, as described below. Then these two groups are used to reset the algorithm in the second stage. The resulting adaptive procedure seems to converge with a few iterations to the true parameter distributions.

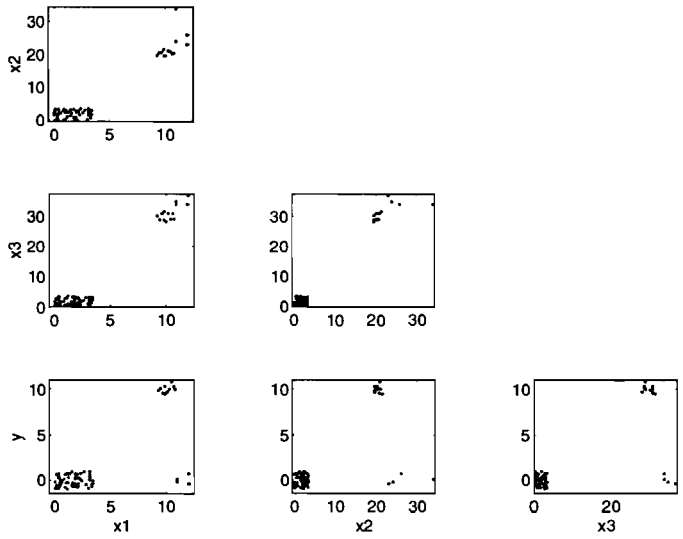


Figure 1: Matrix plot of the Hawkins, Bradu and Kass data.

3.1 First selection of the initial values

The procedure is first initialized by given value zero to the classification variables of data in a set $\boldsymbol{S}_0 = (\boldsymbol{y}_0, \boldsymbol{X}_0)$ and value one otherwise. The set $\boldsymbol{S}_0$ is chosen as a subsample of size n_0 such that the probability of containing more than one outlier is very low. Then we guarantee that: (1) if $\boldsymbol{S}_0$ has no outliers we will obtain unbiased parameter estimates that will lead to the identification of the outliers in the next Gibbs sampler iteration; (2) if $\boldsymbol{S}_0$ has just one outlier, although it can produce biased estimation, obviously, it can not produce masking. In such case, this isolated outlier can be easily detected and then rejected by individual standard diagnostic procedures, as the Bayes factor that a particular observation comes from the alternative distribution against all the data come from the central distribution. The weight of evidence can be done by using Jeffreys (1961, Appen. B) scale of evidence. The Bayes factor is inversely proportional to the conditional predictive ordinate $p(y_j \mid \boldsymbol{y}_{0_{(j)}})$ given by (2.2) and it is a monotonic function of the studentized residuals given by

$$t_j = \frac{y_j - \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_0}{s_{0_{(j)}}(1 - h_{0_j})^{1/2}} \qquad j \in \boldsymbol{S}_0, \tag{3.1}$$

n_0	2	3	4	5	6	7
P_{n_0}	0.990	0.971	0.946	0.915	0.879	0.840

Table 1: Probability of at most one outlier in any set of size n_0 in the Hawkins, Bradu and Kass data with $\alpha_0 = 0.1$.

where $\hat{\boldsymbol{\beta}}_0 = (\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'\boldsymbol{y}_0$ is the least square estimate for the subsample $\boldsymbol{S}_0$, $h_{0_j} = \boldsymbol{x}_j'(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{x}_j$ is the leverage and $s_{0_{(j)}}$ is the sample standard deviation when the jth case is excluded and the sample is $\boldsymbol{S}_0$. We can also test the single outlier by the studentized residuals which have a t distribution with $n_0 - p - 1$ degrees of freedom under the null hypothesis. The global significant level test can be chosen by Bonferroni inequality, that is $\alpha_{\mathrm{T}} = \alpha_{\mathrm{I}}/n_0$.

Let P_{n_0} be the probability that the set $\boldsymbol{S}_0$ contains at most one outlier. As α is the prior probability of each observation being an outlier, then $n(1-\alpha)$ observations in the sample are expected to be good and $n\alpha$ to be outliers. The probability P_{n_0} is computed by the following expression

$$P_{n_0} = \binom{\bar{n}_\alpha}{n_0}\binom{n}{n_0}^{-1} + \binom{\bar{n}_\alpha}{n_0 - 1}\binom{n_\alpha}{1}\binom{n}{n_0}^{-1}, \tag{3.2}$$

where n_α is the nearest integer to $n\alpha_0$ (in case of tie, it is the higher one) and $\bar{n}_\alpha = n - n_\alpha$. Note that α_0 is the expectation of the prior distribution for the parameter α. For instance, Table 1 presents the probabilities P_{n_0} for the artificial data proposed by Hawkins, Bradu and Kass (1984) with $\alpha_0 = 0.1$. Out of the 75 observations in four dimensions data from 1 to 10 are high leverage outliers (see Figure 1). From this table we obtain that if we consider as initial conditions that only three observations come from the central distribution —and we select them randomly—, we expect that this set of size 3 is outlier free in 971 cases out of 1,000 sequences used for the final estimation.

The decision about the size of $\boldsymbol{S}_0$ will be a trade off between sensitivity, that requires the selection of few data points as good data, and power, that depends on having enough data points to estimate the parameters. In any case, we need to take

at least an *elemental set* (Hawkins *et al.*, 1984), that is any set of size p.

3.2 Second selection of the initial values: the Covariance Matrix

The procedure to select the initial conditions in the first stage cannot guarantee that $\boldsymbol{S}_0$ is outlier free. If the initial set $\boldsymbol{S}_0$ contains high leverage outliers, the probability of being outlier will be low for masked outliers and high for swamped good data. The probability of identifying all the outliers will be equal to the probability of non outliers in $\boldsymbol{S}_0$, that is unknown. However, we have seen that the classification variables for groups of masked outliers or swamped good data will have similar behaviour when the series stabilize. Therefore, the covariance matrix of the vector $\boldsymbol{\delta}^{(S)}$ includes information about the dependency among the classification variables that can be useful to identify groups of similar effects. We expect that observations which mask or swamp each other have a large covariance in absolute value, whereas the covariance between outliers and good data points and among good data points will be small. This suggests to estimate the posterior covariance matrix of $\boldsymbol{\delta}^{(S)}$ and to search for sets of points with large covariances in absolute value. These sets are expected to correspond to either masked outliers or swamped good data.

Let $\boldsymbol{C}$ be the Covariance Matrix of the $\boldsymbol{\delta}^{(S)}$ binary variables. Its (i,j) element is

$$c_{ij} = P(\delta_i^{(S)} = 1, \delta_j^{(S)} = 1 \mid \boldsymbol{y}) - P(\delta_i^{(S)} = 1 \mid \boldsymbol{y}) \cdot P(\delta_j^{(S)} = 1 \mid \boldsymbol{y}),$$

and c_{ij} can be estimated by computing the probabilities after S iterations of R parallel replications of the Gibbs sampler. The estimate will be

$$\hat{c}_{ij} = \hat{p}_{ij_R}^{(S)} - \hat{p}_{i_R}^{(S)} \cdot \hat{p}_{j_R}^{(S)},$$

where $\hat{p}_{i_R}^{(S)}$, estimate of $P(\delta_i^{(S)} = 1 \mid \boldsymbol{y})$, is given by (2.6) and $\hat{p}_{ij_R}^{(S)}$, estimate of $P(\delta_i^{(S)} = 1, \delta_j^{(S)} = 1 \mid \boldsymbol{y})$, is given by

$$\hat{p}_{ij_R}^{(S)} = \frac{1}{R} \sum_{r=1}^{R} \delta_{ir}^{(S)} \, \delta_{jr}^{(S)}.$$

This Covariance Matrix is related to the one used by Peña and Tiao (1992) who proposed a probabilistic interaction matrix for computing the curves BROC and SE-BROC. They did not use marginal probabilities, as we do here, but joint probabilities

that: (1) one observation is an outlier and all the others come from the central distribution and (2) two observations are outliers and all the others come from the central distribution.

As we are searching for sets of observations with similar dependency structure it is natural to try to identify these sets by studying the eigenstructure of the matrix $\hat{\boldsymbol{C}}$. Also Peña and Yohai (1995) have shown, in a different context, that outliers can be identified by looking at the eigenstructure of their Influence Matrix. In order to study the eigenvalues of the matrix $\hat{\boldsymbol{C}}$, let us call $\boldsymbol{D}$ to the data matrix for the classification variables after S iterations. This matrix is

$$\boldsymbol{D} = (\boldsymbol{\delta}_1^{(S)}, \ldots, \boldsymbol{\delta}_R^{(S)})', \tag{3.3}$$

where the columns in (3.3) are random samples of each classification variable δ_i at iteration S. Then the matrix $\hat{\boldsymbol{C}}$ may be written as

$$\hat{\boldsymbol{C}} = \frac{1}{R}\boldsymbol{D}'\boldsymbol{D} - \frac{1}{R^2}\boldsymbol{D}'\mathbf{1}_R\mathbf{1}_R'\boldsymbol{D},$$

and the eigenvectors associated to the non null eigenvalues of $\hat{\boldsymbol{C}}$ will be the coefficients of the principal components of $\boldsymbol{D}$.

Let us consider the limit case in which there is only one group of outliers. Then we can obtain the expected behaviour of the eigenvalues and eigenvectors of the matrix $\hat{\boldsymbol{C}}$. Let us call $\boldsymbol{d}_i$ to the ith column vector of $\boldsymbol{D}$ and, without loss of generality, let us assume that the set of outliers corresponds to the last columns of the matrix $\boldsymbol{D}$. In addition, let us call H to the set of swamped data (that may be void) and G to the set of not swamped good data. Let us assume that the sizes of these sets are n_I, n_H and n_G, respectively ($n = n_G + n_H + n_I$), and that the swamped data correspond with the columns before the n_I outliers.

Suppose that the series of outlier probabilities are stable at iteration S and let us call $J_1^{(S)}, \ldots, J_R^{(S)}$ to the sets that the Gibbs sampler identifies as outliers in each run. By the initial condition dependence property, $J_r^{(S)}$ is equal to I when the initial set $\boldsymbol{S}_0$ is outlier free. Let us call q to this probability. Then $\boldsymbol{S}_0$ will be outlier free in $Q = qR$ of these sets. Let $\bar{Q} = R - Q$. In order to analyze the expected behaviour of the elements of the vectors $\boldsymbol{d}_i$, let us assume, without loss of generality, that the first Q

runs correspond to the Q outlier-free initial conditions. We distinguish the following types of column vectors in $\boldsymbol{D}$:

(a) Columns which correspond to the not swamped good data are of the form

$$\boldsymbol{d}_j = \begin{pmatrix} \mathbf{0}_{\mathrm{Q}} \\ \boldsymbol{g}_j \end{pmatrix} \qquad j = 1, \ldots, n_{\mathrm{G}},$$

where $\mathbf{0}_{\mathrm{Q}}$ is a $Q \times 1$ null vector by the initial condition dependence property, and the vector $\boldsymbol{g}_j = (g_{1j}, \ldots, g_{\bar{\mathrm{Q}}j})'$ may contains a few non null elements because the outlier probability for good data is small, but not zero. We may suppose there are not important differences between these columns in the proportion of ones (misspecifications), that is bounded by some small value π, such that

$$\frac{1}{R} \sum_{i=1}^{\bar{\mathrm{Q}}} g_{ij} \leq \pi \quad \text{for all } j = 1, \ldots, n_{\mathrm{G}}. \tag{3.4}$$

(b) Columns which correspond with swamped good data, due to some not identified outliers are of the form

$$\boldsymbol{d}_j = \begin{pmatrix} \mathbf{0}_{\mathrm{Q}} \\ \mathbf{1}_{\bar{\mathrm{Q}}} \end{pmatrix} \qquad j = n_{\mathrm{G}} + 1, \ldots, n_{\mathrm{G}} + n_{\mathrm{H}},$$

where $\mathbf{1}_{\bar{\mathrm{Q}}}$ is a $\bar{Q} \times 1$ unit vector.

(c) Columns which correspond with data in the group of outliers are of the form

$$\boldsymbol{d}_j = \begin{pmatrix} \mathbf{1}_{\mathrm{Q}} \\ \boldsymbol{g}_j \end{pmatrix} \qquad j = n - n_{\mathrm{I}}, \ldots, n, \tag{3.5}$$

where

$$0 \leq \sum_{j=n-n_{\mathrm{I}}}^{n} g_{ij} < n_{\mathrm{I}} \qquad i = 1, \ldots, \bar{Q}.$$

The number of unity elements in $\boldsymbol{g}_j$ depends on the degree of masking. The two extreme cases are: (1) the outliers in I are isolated outliers, that implies $\boldsymbol{g}_j = \mathbf{1}_{\bar{\mathrm{Q}}}$; and (2) the data in I are identical high leverage outliers, that implies $\boldsymbol{g}_j = \mathbf{0}_{\bar{\mathrm{Q}}}$. Let us consider this last case in which Gibbs sampling has failed completely. Then the column vectors (3.5) are

$$\boldsymbol{d}_j = \begin{pmatrix} \mathbf{1}_{\mathrm{Q}} \\ \mathbf{0}_{\bar{\mathrm{Q}}} \end{pmatrix} \qquad j = n - n_{\mathrm{I}}, \ldots, n.$$

By (a)-(c) the matrix $\boldsymbol{D}$ may be expressed as a block matrix

$$\boldsymbol{D} = \begin{pmatrix} \mathbf{0} & \vdots & \mathbf{0} & \vdots & \mathbf{1}_{\bar{\mathrm{Q}}}\mathbf{1}'_{n_{\mathrm{I}}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{G} & \vdots & \mathbf{1}_{\bar{\mathrm{Q}}}\mathbf{1}'_{n_{\mathrm{II}}} & \vdots & \mathbf{0} \end{pmatrix},$$

where $\boldsymbol{G} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{n_{\mathrm{G}}})$ is a matrix $\bar{Q} \times n_{\mathrm{G}}$. The Covariance Matrix $\hat{\boldsymbol{C}}$ can be written as

$$\hat{\boldsymbol{C}} = \begin{pmatrix} \frac{1}{R}\boldsymbol{G}'\boldsymbol{G} - \frac{1}{R^2}\boldsymbol{G}'\mathbf{1}_{\bar{\mathrm{Q}}}\mathbf{1}'_{\bar{\mathrm{Q}}}\boldsymbol{G} & \vdots & \frac{Q}{R^2}\boldsymbol{G}'\mathbf{1}_{\bar{\mathrm{Q}}}\mathbf{1}'_{n_{\mathrm{H}}} & -\frac{Q}{R^2}\boldsymbol{G}'\mathbf{1}_{\bar{\mathrm{Q}}}\mathbf{1}'_{n_{\mathrm{I}}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{Q}{R^2}\mathbf{1}_{n_{\mathrm{II}}}\mathbf{1}'_{\bar{\mathrm{Q}}}\boldsymbol{G} & \vdots & \frac{Q\bar{Q}}{R^2}\mathbf{1}_{n_{\mathrm{II}}}\mathbf{1}'_{n_{\mathrm{H}}} & -\frac{Q\bar{Q}}{R^2}\mathbf{1}_{n_{\mathrm{II}}}\mathbf{1}'_{n_{\mathrm{I}}} \\ -\frac{Q}{R^2}\mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{\bar{\mathrm{Q}}}\boldsymbol{G} & \vdots & -\frac{Q\bar{Q}}{R^2}\mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{n_{\mathrm{H}}} & \frac{Q\bar{Q}}{R^2}\mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{n_{\mathrm{I}}} \end{pmatrix}.$$

Assuming that π is small, this matrix can be approximated by

$$\hat{\boldsymbol{C}} = \begin{pmatrix} \frac{1}{R}\boldsymbol{G}'\boldsymbol{G} & \vdots & \mathbf{0} \\ \cdots & \cdots & \cdots \\ \mathbf{0} & \vdots & \hat{\boldsymbol{C}}_{22} \end{pmatrix},$$

where $\hat{\boldsymbol{C}}_{22}$ is the $(n_{\mathrm{II}} + n_{\mathrm{I}}) \times (n_{\mathrm{II}} + n_{\mathrm{I}})$ matrix

$$\hat{\boldsymbol{C}}_{22} = \frac{Q\bar{Q}}{R^2}\begin{pmatrix} \mathbf{1}_{n_{\mathrm{H}}}\mathbf{1}'_{n_{\mathrm{II}}} & \vdots & -\mathbf{1}_{n_{\mathrm{II}}}\mathbf{1}'_{n_{\mathrm{I}}} \\ \cdots & \cdots & \cdots \\ -\mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{n_{\mathrm{II}}} & \vdots & \mathbf{1}_{n_{\mathrm{I}}}\mathbf{1}'_{n_{\mathrm{I}}} \end{pmatrix}.$$

The eigenvalues of $\hat{\boldsymbol{C}}$ are the eigenvalues of the matrices $\boldsymbol{G}'\boldsymbol{G}/R$ and $\hat{\boldsymbol{C}}_{22}$. By equation (3.4) the eigenvalues of $\boldsymbol{G}'\boldsymbol{G}/R$ verify

$$\sum_{i=j}^{n_{\mathrm{G}}} \lambda_j = \mathrm{tr}\left(\frac{1}{R}\boldsymbol{G}'\boldsymbol{G}\right) = \frac{1}{R}\sum_{j=1}^{n_{\mathrm{G}}}\sum_{i=1}^{\bar{Q}} g_{ij}^2 \leq \pi\, n_{\mathrm{G}}.$$

The matrix $\hat{\boldsymbol{C}}_{22}$ has only one non null eigenvalue, given by

$$\lambda_{\mathrm{I}} = q(1-q)(n_{\mathrm{H}} + n_{\mathrm{I}}). \tag{3.6}$$

Then the matrix $\hat{\boldsymbol{C}}$ has an eigenvalue λ_{I} and n_{G} additional eigenvalues such that their sum is less or equal than πn_{G}, where π is very close to zero. In addition, $\boldsymbol{v}_a = (\mathbf{0}'_{n_{\mathrm{G}}}, a\mathbf{1}'_{n_{\mathrm{H}}}, -a\mathbf{1}'_{n_{\mathrm{I}}})'$ is an eigenvector of the matrix $\hat{\boldsymbol{C}}$ associated with λ_{I}, for all non null values of a.

The λ_{I} eigenvalue, given by (3.6) in the case of only one group of outliers, may be close to zero (the group is unidentified) when the probability q of outlier-free initial

conditions is close to zero or one. A value of q close to zero corresponds to the strong contamination case and a large size of $\boldsymbol{S}_0$. We avoid this problem with the proposed procedure by selecting a small initial set $\boldsymbol{S}_0$ as it was described before. On the other hand, a value of q close to one corresponds to the case in which there is not outliers in the sample, or only very few and small size of $\boldsymbol{S}_0$. In this case, the outliers will not be masked and they can be directly detected by the Gibbs sampling algorithm. The interesting case is when $0 < q < 1$ and n_{I} (and may be n_{II}) is large, that corresponds to the most difficult case in which outliers not only are not identified in most run, but also they are producing swamping. Then λ_{I} will be relatively large and the eigenvector linked to this eigenvalue will indicate correctly the masked and swamped data. The observations having relatively large coefficient (in absolute value) on the eigenvector $\boldsymbol{v}_a$ are potentially outlier candidates, and we may split the data into two subsets: (1) the set that contains the observations with non null coefficients on the eigenvector $\boldsymbol{v}_a$ or with high individual probability $\hat{p}_i^{(s)}$; and (2) the set of the remainder observations. We call to the first set the *potential outlier set* (PO).

For instance, Table 2 shows the Covariance Matrix for the data provided by Hawkins, Bradu and Kass (1984) and showed in Figure 1. It is a well-known example of data with high leverage outliers where the traditional outlier identification procedures are not able to identify the outliers and, even worse, observations 11 to 14 are good data identified wrongly as outliers. Justel and Peña (1996) show that Gibbs sampling fail with this data set. The ten outliers are not identified and the Gibbs sampling suffers the same problems as traditional methods for outlier detection. The Gibbs sampling is started with a set $\boldsymbol{S}_0$ of four observations considered as good data point, therefore the probability of non outliers in $\boldsymbol{S}_0$ is

$$q = \binom{10}{0}\binom{65}{4} \Big/ \binom{75}{4} = 0.557.$$

The largest eigenvalues are shown in Table 3, and the components of the eigenvector associated with the highest eigenvalue are shown in Figure 2. We shall include in PO the observations 1 to 14. For this data, the matrix $\hat{\boldsymbol{C}}$ was built with the estimated probabilities after 500 iterations. Note that here $q = 0.557$, $n_{\text{I}} = 10$, $n_{\text{H}} = 4$, and therefore the expected value of the largest eigenvalue is, according to (3.6) equal to 3.45 that is very similar to the real observed value.

2	.22															
3	.23	.22														
4	.22	.22	.22													
5	.22	.22	.23	.22												
6	.23	.22	.23	.22	.23											
7	.22	.22	.22	.21	.22	.22										
8	.22	.22	.22	.22	.22	.22	.21									
9	.22	.22	.22	.22	.22	.22	.22	.22								
10	.23	.22	.23	.22	.23	.23	.22	.22	.22							
11	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22						
12	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	.22					
13	-.22	-.22	-.22	-.22	-.22	-.22	-.21	-.22	-.22	-.22	.22	.22				
14	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	-.22	.22	.22	.22			
15	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Table 2: Covariance Matrix with Hawkins, Bradu and Kass data. Only values greater than .01 and less than $-.01$ are printed.

Component number	Eigenvalue	Variance	Cum. percentage of variance
1	3.4297	0.7849	78.49
2	0.0391	0.0089	79.38
3	0.0338	0.0077	80.15

Table 3: Covariance Matrix eigenvalues for the Hawkins, Bradu and Kass data.

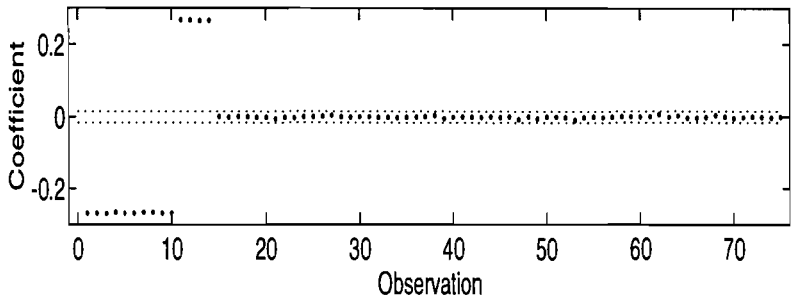


Figure 2: Coefficients of the eigenvector associated with the eigenvalue λ_1 of the Covariance Matrix with Hawkins, Bradu and Kass data.

When the sample data contains several sets of outliers they can produce p different independent effects in $\mathbf{R}^p$. Therefore, the maximum number of eigenvalues to scrutinized is p. A straightforward generalization of the previous analysis shows that this independent effects will appear in p eigenvectors of the estimated Covariance Matrix $\hat{C}$. This result is the basis of the procedure presented in the next section.

3.3 Algorithm for sampling posterior probabilities

The method for the first selection of the initial values, together with the information provided by the Covariance Matrix, allows to split the data into two sets PO and $\overline{\text{PO}}$. If the Gibbs sampler is initialized giving value 1 to the classification variables in PO for each sequence, after a few iterations the classification variables obtained are a sample from the posterior distribution. Inference from this sample allows us to identify the outliers. Accordingly, we suggest an Adaptive Gibbs Sampling Algorithm following two stages:

Stage 1: Run the Gibbs sampling until the series of posterior outlier probabilities become stable. The initial conditions for each sequence are selected as follows:

i. Let n_0 be the maximum integer such that the probability (3.2) of finding at most one outlier in any data subset of size n_0 is greater than c_1. Then select $m = \max\{n_0, p\}$ random numbers $i_1, \ldots, i_m$ among $1, \ldots, n$.

ii. Build the initial set $\boldsymbol{S}_0 = \{(y_{i_1}, \boldsymbol{x}_{i_1}), \ldots, (y_{i_m}, \boldsymbol{x}_{i_m})\}$. If $m > p$, compute the studentized residuals $t_{i_1}, \ldots, t_{i_m}$ given by the expression (3.1).

iii. When $m = p$, the initial classification variables are:

$$\delta_j^{(0)} = \begin{cases} 0 & \text{if } j = i_1, \ldots, i_m \\ 1 & \text{otherwise.} \end{cases}$$

When $m > p$, the initial classification variables are:

$$\delta_j^{(0)} = 1 \qquad \text{for } j \neq i_1, \ldots, i_m$$

$$\delta_j^{(0)} = \begin{cases} 0 & \text{if } t_j < t_{m-p-1,\alpha_I/n_0} \quad (\textit{Student } t) \\ 1 & \text{otherwise} \end{cases} \qquad \text{for } j = i_1, \ldots, i_m.$$

iv. If $n - \sum \delta_j < p$ or the resulting matrix with the rows which correspond to $\delta_j^{(0)} = 0$ is not positive define, execute again steps *i-iii.* Otherwise, $\beta^{(0)} = (\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{(0)^{-1}}\boldsymbol{y}$, where $\boldsymbol{V}^{(0)}$ is a diagonal matrix with $v_{jj}^{(0)} = 1 + \delta_j^{(0)}(k-1)$.

With the values obtained in the last iteration compute the Covariance Matrix $\hat{\boldsymbol{C}}$ and the largest c_2 eigenvalues and associated eigenvectors $(v_1, v_2, \ldots)$. Split the sample into two sets PO and $\overline{\text{PO}}$ as follows:

a. If $\hat{p}_i^{(s)} > 0.5$, then $(y_j, \boldsymbol{x}_j') \in$ PO.

b. For $i = 1, \ldots, c_2$ and $j = 1, \ldots, n$, compute $m_i = \text{median} \mid v_{ij} \mid /0.6475$. If $\mid v_{ij} \mid > c_3\, m_i$, then $(y_j, \boldsymbol{x}_j') \in$ PO.

c. If $(y_j, \boldsymbol{x}_j') \notin$ PO, then it is in $\overline{\text{PO}}$.

Stage 2. Reset the algorithm and run the Gibbs sampling until the series of posterior outlier probabilities become stable. The initial conditions for each sequence are:

1. $\delta_j^{(0)} = 1$ if $(y_j, \boldsymbol{x}_j') \in$ PO, and $\delta_j^{(0)} = 0$ otherwise.
2. $\boldsymbol{\beta}^{(0)} = (\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{y}$, where $\boldsymbol{V}^{(0)}$ is a diagonal matrix with $v_{jj}^{(0)} = 1 + \delta_j^{(0)}(k-1)$.

The interpretation of Stage 1 is clear: to obtain a set $\boldsymbol{S}_0$ with a small probability of containing outliers, then we split the sample using the information from the Covariance

Matrix. The points with large coordinates on the eigenvectors are obtained by using a robust measure to deviations from zero. Finally, in Stage 2 the algorithm is reset and the procedure is run again. The procedure ends when the final series of outlier probabilities become stable.

The bounds c_1 and c_2 and the constant c_3 must be chosen. The criterion for c_1 was discussed in section 3.1 and we suggest values around 0.9 in order to consider both sensitivity and power. We suggest to choose the minimum value of $(p, c_2^\star)$, where $c_2^\star$ is the number of eigenvalues greater than five times a robust dispersion measure of the eigenvalues λ_i of $\hat{\boldsymbol{C}}$, that can be $median(\lambda_i)/0.6475$. The constant c_3 is used to determine the significative non null coordinates in the eigenvectors and, therefore, the outlier candidates. We use again a robust measure of the dispersion around zero, that is the expected value for the good data. The number of parallel sequences depends on the asymptotic properties of the estimates. Finally, the number of iterations needed to achieve the series stabilization in both stages may be decided by the methods for monitoring convergence proposed by Gelman and Rubin (1992) or Robert (1994), among others. We suggest an easier procedure that in this particular application of the Gibbs sampling seems to work well. The Gibbs sampler is run until the iteration S, such that, given $\epsilon > 0$, $\mid \hat{p}_{i_{\mathrm{R}}}^{(\mathrm{S}+1)} - \hat{p}_{i_{\mathrm{R}}}^{(\mathrm{S})} \mid < \epsilon$ for all $i = 1, \ldots, n$. Finally, in the Stage 2 the initial conditions are always the same and it is possible to run only one sequence to reduce the computational effort.

4 PROCEDURE PERFORMANCE

We compare the performance of the new method with the two versions of the procedure to identify multiple outliers by Hadi and Simonoff (1993) and with the one by Peña and Yohai (1995). In both procedures the outliers are the observations with large studentized residuals in a regression computed from a subsample that is supposed to be outlier free. Therefore, some of the residuals we will display in the tables are the out-of-sample residuals (note that we will not differentiate these points). We present the results of the first method suggested by Hadi and Simonoff (1993). The performance of the second one is similar to the first in all the three examples analyzed.

In our application we choose $c_1 = 0.95$, $c_3 = 5$ and the individual significance level $\alpha_I = 0.05$. The number c_2 of eigenvectors to be examined is decided by the method explained before. The Gibbs sampler is always run 300 sequences and the number of iteration is decided with $\epsilon = 0.002$. In all the examples $k = 10$, $\alpha_0 = 0.2$ and $\gamma_1 + \gamma_2 = n$, that imply $E(\alpha \mid \delta) = 1/2E(\alpha) + 1/2\bar{\delta}$. Then $\boldsymbol{\beta}^{(0)}$ is the generalized least square estimate, $\boldsymbol{\beta}^{(0)} = (\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{(0)^{-1}}\boldsymbol{y}$, where $\boldsymbol{V}^{(0)}$ is a diagonal matrix with elements $k^2\sigma^2$ if $\delta_i^{(0)} = 1$ and σ^2 otherwise. It is not necessary to specify the initial value for the variance because it is the first parameter computed in the iterations and for α because only depends on $\boldsymbol{\delta}$. The last iteration of each performance is used to estimate the posterior outlier probabilities.

4.1 Stars data

The scatter plot displayed in Figure 3 represents the Hertzsprung-Russell diagram of the star cluster CYG OB1 from Rousseeuw and Leroy (1987). The data correspond to 47 stars in the direction of Cygnus and the variables are the logarithm of the effective temperature at the surface of the star (x) and the logarithm of the light intensity (y). There are four outliers which correspond to giant stars in the data points 11, 20, 30 and 34. The other observations more distant to the cluster are the data points 7, 9 and 18. The studentized residuals obtained with the procedures by Hadi and Simonoff (1993), as well as the procedure by Peña and Yohai (1995), are shown in Table 4, columns 1–3. The three methods are successful in identifying the outliers.

The posterior outlier probabilities after the first run of the Gibbs sampling are represented by a bar in Figure 5(a). These probabilities identify the group of outliers since their outlier probabilities are greater than 0.5. The Gibbs sampling starts in this Stage 1 with an initial set $\boldsymbol{S}_0$ of size three, and the eigenvalues of the Covariance Matrix that must be examined are $p = 2$ ($\lambda_1 = 0.97$ and $\lambda_2 = 0.21$). The two eigenvectors associated with the largest eigenvalues are showed in Figure 4. The points are the coordinates of each data in the eigenvector and the dotted lines are the zero confidence bands. In both eigenvalues the outliers are outside the confidence bands and in the second eigenvalue the coordinates corresponding to the data 7, 9 and 18 are also non null and with opposite signs to the outliers. The coordinates of the good data points

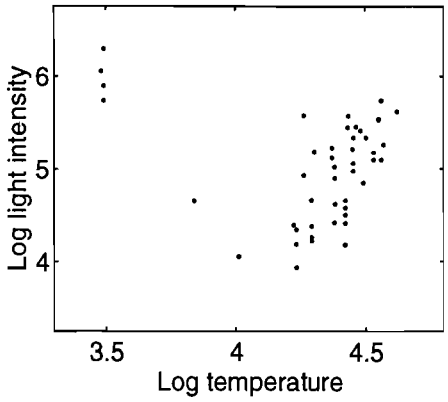


Figure 3: Hertzsprung-Russell diagram of the star cluster CYG OB1.

5, 14 and 40 are in the limit of the bands and these are also considered as potential outliers. Therefore the PO set includes the outliers and this information is used to select the initial conditions for the Gibbs sampling in the Stage 2. It can be seen in Figure 5(b) that the four giant stars are clearly confirmed as outliers with probabilities greater than 0.5.

4.2 Hawkins, Bradu and Kass data

In the second example, the procedure is applied to the Hawkins, Bradu and Kass data discussed in sections 2 and 3. The observations 1 to 10 are outliers which swamp the good data 11 to 14. In this data set the procedures by Hadi and Simonoff (1993) fail due to the high leverage of the outliers, whereas the one by Peña Yohai (1995) is successful in identifying the outliers. It can be seen in Table 5 that the largest residuals provided by the Hadi and Simonoff (1993) procedures correspond to the good data and that the outliers are masked.

The initial conditions in the Stage 1 include a set of four observations considered as good, that is the size of the elemental set. The number of eigenvalues of the Covariance Matrix to be examined by the algorithm is one, and the associated eigenvector is showed in Figure 2. In this example the estimates of the individual probabilities, showed in Figure 6(a), and the eigenstructure of the Covariance Matrix, discussed in

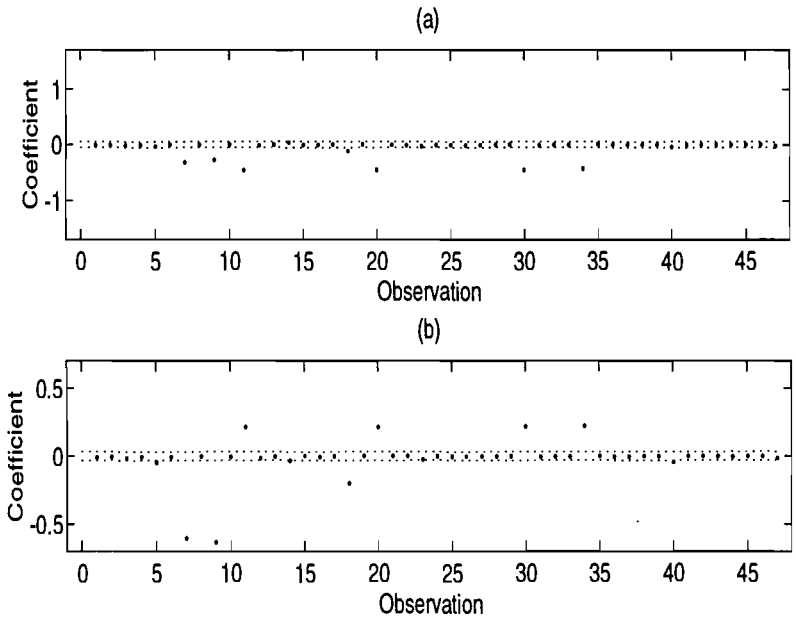


Figure 4: Coefficients of the eigenvectors associated with the eigenvalues λ_1 (in (a)) and λ_2 (in (b)) of the Covariance Matrix with the Stars data.

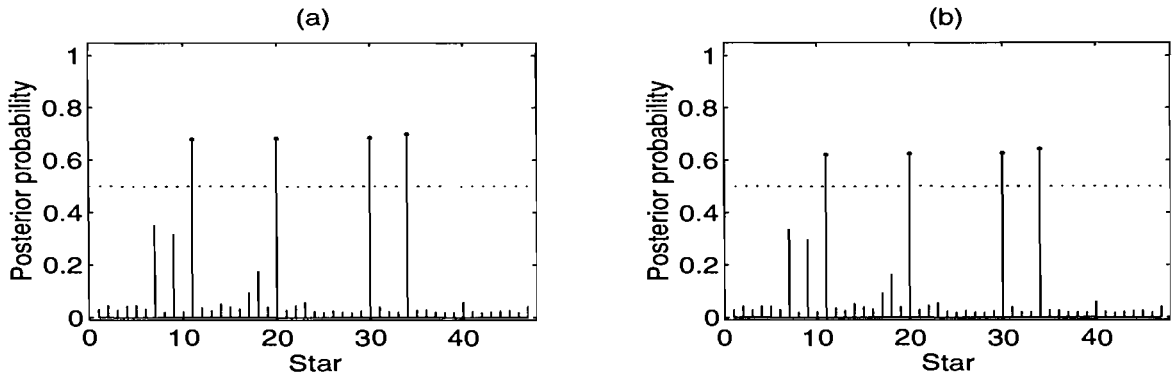


Figure 5: Results of the Gibbs sampler with the Stars data: (a) probabilities of each data point to be outlier in the Stage 1; (b) posterior outlier probabilities in the Stage 2.

Data	Student. res. HS	PY	Prob. AGSA	Data	Student. res. HS	PY	Prob. AGSA
1	0.264	0.861	0.028	26	-2.686	-0.829	0.025
2	1.670	1.195	0.042	27	-1.930	-0.159	0.019
3	-0.578	0.677	0.027	28	-1.403	-0.019	0.017
4	1.670	1.195	0.042	29	-2.730	-0.483	0.024
5	0.480	1.133	0.045	**30**	**5.607**	**5.328**	**0.627**
6	0.862	0.982	0.030	31	-3.576	-1.238	0.039
7	0.868	2.726	0.337	32	-1.485	-0.447	0.021
8	-0.707	-0.067	0.019	33	-0.246	0.429	0.019
9	2.386	2.437	0.296	**34**	**6.310**	**5.750**	**0.643**
10	-0.272	0.582	0.022	35	-3.018	-0.662	0.026
11	**4.673**	**4.745**	**0.620**	36	0.740	0.570	0.026
12	0.986	1.113	0.036	37	-1.302	-0.289	0.019
13	0.553	0.775	0.025	38	-0.246	0.422	0.019
14	-2.845	-0.273	0.051	39	-0.910	-0.087	0.018
15	-3.765	-1.180	0.039	40	1.570	1.430	0.056
16	-3.059	-1.035	0.031	41	-2.644	-0.722	0.023
17	-4.821	-1.727	0.092	42	-1.024	0.022	0.017
18	-4.927	-2.126	0.162	43	0.048	0.469	0.020
19	-3.739	-1.077	0.039	44	0.336	0.726	0.024
20	**5.141**	**5.032**	**0.623**	45	0.741	0.723	0.025
21	-3.214	-0.868	0.028	46	-1.414	-0.177	0.018
22	-3.948	-1.286	0.044	47	-3.433	-1.244	0.040
23	-3.807	-1.456	0.055				
24	-2.182	-0.713	0.023				
25	-0.818	0.279	0.018				

Table 4: Results with the procedures by Hadi and Simonoff (HS), Peña and Yohai (PY) and the Adaptive Gibbs Sampling Algorithm (AGSA) with the Stars data.

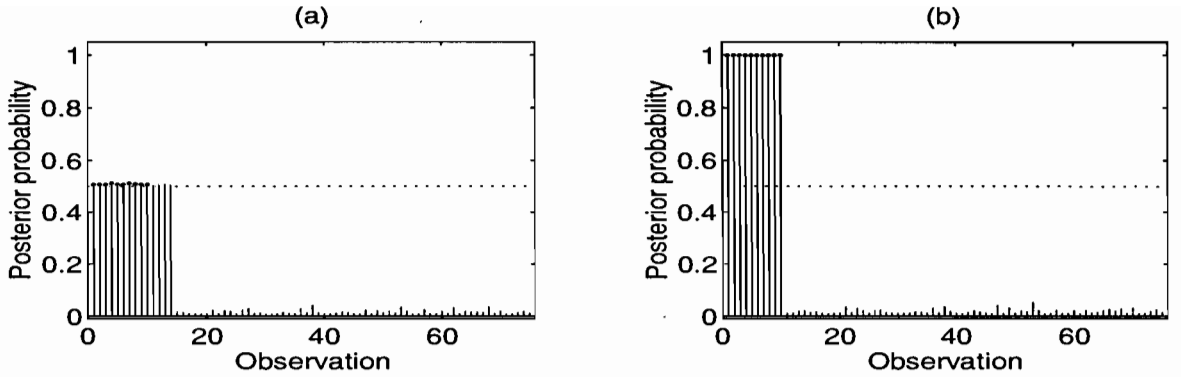


Figure 6: Results of the Gibbs sampler with Hawkins, Bradu and Kass data: (a) probabilities of each data point to be outlier in the Stage 1; (b) posterior outlier probabilities in the Stage 2.

section 3, lead to the same conclusion: the group of potential outliers PO includes the observations 1 to 14, that are the masked outliers and the swamped good data. In the Stage 2 these data points are considered outliers in the initial conditions and the outliers are correctly identified with probability equal to one (see Figure 6(b) for the posterior outlier probabilities). Note that the probabilities showed in the last column of the Table 5 are very low for the four previously swamped data.

4.3 Rousseeuw data

This set of simulated data from Rousseeuw (1984) is the most interesting because it shows the high breakdown point of the procedure based on the Gibbs sampler. The contamination is 40 per cent and the procedures by Hadi and Simonoff (1993) and Peña and Yohai (1995), are not able to unmask the outliers (see Table 7). The data are generated in two groups that can be seen in the scatter plot of Figure 7. The first group, that is on the right of the plot, follows an spherical distribution, whereas the second group follows the linear model $y_i = 2 + x_i + u_i$ with error standard deviation 0.2. Out of the 50 data points, 20 are high leverage outliers and 30 good observations (see Table 6 for the numerical values).

The usual diagnostic procedures identify as outliers the observations 32 and 33 that

Data	Student. res. HS	PY	Prob. AGSA
1	1.0762	5.3525	1.0000
2	2.2188	5.4420	1.0000
3	0.1100	5.3188	1.0000
4	-1.5237	4.8893	1.0000
5	-0.1409	5.1448	1.0000
6	0.7106	5.3135	1.0000
7	2.9565	5.6465	1.0000
8	2.2196	5.5893	1.0000
9	-0.6850	5.0402	1.0000
10	0.8538	5.3079	1.0000
11	-26.6269	0.9464	0.0117
12	-28.7513	0.9020	0.0117
13	-25.1989	0.6873	0.0185
14	-11.8374	0.8719	0.0194

Table 5: Results with the procedures by Hadi and Simonoff (HS), Peña and Yohai (PY) and the Adaptive Gibbs Sampling Algorithm (AGSA) with the Hawkins, Bradu and Kass data.

are good data with large least square residuals. The solid line in the Figure 7 is the least square estimate of the regression line. Also the standard Gibbs sampler does not identify the outliers as Justel and Peña (1996) showed. However, the Adaptive Gibbs Sampling Algorithm proposed in this paper works very well. Starting with a set of four good observations, the outlier probabilities in the Stage 1 for the 20 outliers are low (see Figure 9(a)), but the Covariance Matrix has two non null eigenvalues $\lambda_1 = 0.53$ and $\lambda_2 = 0.31$. The coordinates of the associated eigenvectors are showed in Figure 8. In the first eigenvector the results are as expect: (1) the coordinates are non null for the 20 outliers and the swamped good data; and (2) the signs are opposite for the group of outliers and for the swamped data. Then the PO group includes the 20 outliers and the observations 32 and 33. The posterior outlier probabilities estimated in the second stage (see Figure 9(b)) are such that the outliers are correctly identify in a few iterations and also the swamping effect disappears.

x	7.46	6.90	6.99	6.79	7.01	7.03	7.10	6.97	7.27	6.83
y	1.68	1.90	2.27	2.97	1.89	1.53	2.01	1.51	1.32	1.56
x	6.56	7.22	6.70	7.68	6.80	6.30	6.43	6.69	7.66	7.20
y	2.24	1.05	1.43	2.60	1.61	3.41	2.01	1.77	1.06	2.41
x	2.74	2.24	2.61	1.72	1.23	2.25	1.46	1.88	2.74	2.28
y	5.05	3.84	4.73	4.04	2.89	4.09	3.61	3.94	4.68	3.75
x	2.58	3.71	3.89	1.96	1.01	2.76	2.10	1.59	3.23	1.39
y	4.32	5.88	6.10	3.89	3.04	4.58	4.27	3.66	5.33	3.61
x	1.24	1.71	2.94	1.09	3.29	2.21	2.32	1.27	1.87	2.28
y	3.31	3.38	5.02	2.87	5.14	4.22	4.39	3.03	4.15	4.22

Table 6: Rousseeuw data.

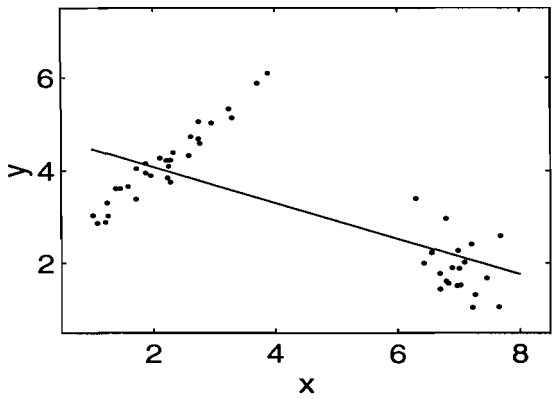


Figure 7: Least square estimate with the Rousseeuw data.

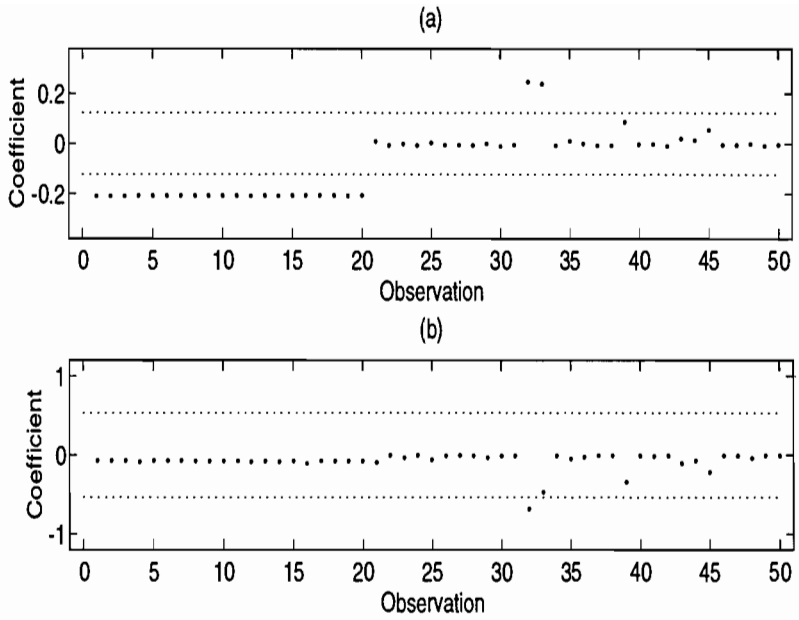


Figure 8: Coefficients of the eigenvectors associated with the eigenvalues λ_1 (in (a)) and λ_2 (in (b)) of the Covariance Matrix with Rousseeuw data.

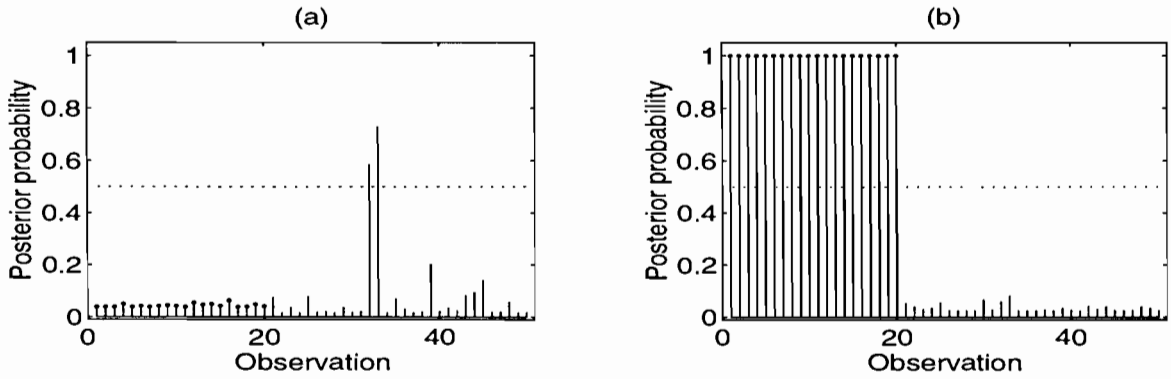


Figure 9: Results of the Gibbs sampler with Rousseeuw data: (a) probabilities of each data point to be outlier in the Stage 1; (b) posterior outlier probabilities in the Stage 2.

Data	Student. res. HS	PY	Prob. AGSA	Data	Student. res. HS	PY	Prob. AGSA
1	0.963	-0.124	1.000	26	0.233	-0.653	0.025
2	0.714	-0.232	1.000	27	-2.916	-1.825	0.025
3	2.211	0.311	1.000	28	-1.092	-1.118	0.024
4	4.407	1.074	1.000	29	3.275	0.440	0.024
5	0.895	-0.168	1.000	30	-1.049	-1.077	0.066
6	-0.478	-0.614	1.000	31	1.778	-0.127	0.027
7	1.576	0.056	1.000	32	9.540	2.868	0.059
8	-0.648	-0.673	1.000	33	10.934	3.420	0.081
9	-0.845	-0.723	1.000	34	-1.131	-1.126	0.024
10	-0.742	-0.712	1.000	35	-5.750	-2.911	0.026
11	1.384	-0.031	1.000	36	2.952	0.329	0.024
12	-1.843	-1.110	1.000	37	0.656	-0.523	0.028
13	-1.499	-0.971	1.000	38	-2.536	-1.675	0.024
14	4.602	1.233	1.000	39	6.590	1.648	0.035
15	-0.614	-0.672	1.000	40	-3.067	-1.888	0.027
16	5.165	1.305	1.000	41	-4.400	-2.386	0.024
17	0.259	-0.403	1.000	42	-3.314	-1.941	0.041
18	-0.190	-0.538	1.000	43	4.924	1.028	0.030
19	-1.105	-0.788	1.000	44	-6.201	-3.064	0.039
20	3.099	0.640	1.000	45	5.993	1.426	0.025
21	4.616	0.912	0.052	46	0.645	-0.520	0.024
22	-0.788	-0.992	0.041	47	1.532	-0.224	0.025
23	3.208	0.412	0.029	48	-5.357	-2.729	0.039
24	-1.023	-1.106	0.035	49	-0.283	-0.848	0.033
25	-5.911	-2.937	0.054	50	0.811	-0.461	0.024

Table 7: Results with the procedures by Hadi and Simonoff (HS), Peña and Yohai (PY) and the Adaptive Gibbs Sampling Algorithm (AGSA) with the Rousseeuw data.

5 CONCLUDING REMARKS

The Bayesian procedure proposed in this paper for outlier detection in linear models combines in a sequential learning procedure the Gibbs sampling with the information from an estimate of the Covariance Matrix of the classification variables. The eigenvectors associated to the non zero eigenvalues of this matrix provide information about which data are outlier candidates. The procedure can be used automatically and includes: (1) a criterion for initial conditions selection without any prior information; and (2) a method to be used for grouping data based on the Covariance Matrix. Its application to some of the most frequently used examples in multiple outlier detection shows that it is able to unmask outliers in samples where other methods fail.

ACKNOWLEDGEMENTS

Both authors would like to acknowledge support for this research provided by DGICYT (Spain), under grant No. PB93–0232.

REFERENCES

ATKINSON, A.C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association,* **89,** 1329–1339.

BOX, G.E.P. & TIAO, C.G. (1968). A Bayesian approach to some outlier problems. *Biometrika,* **55**, 119–129.

BOX, G.E.P. & TIAO, C.G. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Mass: Addison-Wesley.

COOK, R.D. & WEISBERG, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

FREEMAN, P.R. (1980). On the number of outliers in data from a linear model. In *Bayesian Statistics 1,* Ed. J.M. BERNARDO, M.H. DEGROOT, D.V. LINDLEY and A.F.M. SMITH, pp. 349–365. Valencia University Press.

GEISSER, S. (1980). Discussion of a paper by G.E.P. BOX. *Journal of the Royal Statistical Society A,* **143,** 416-417.

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

GELMAN, A. & RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science,* **7,** 457–511.

GEMAN, S. & 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.

HADI, A.S. & SIMONOFF, J.S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association,* **88,** 1264–1272.

HAWKINS, D.M., BRADU, D. & KASS, G.V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics,* **26,** 197–208.

JEFFREYS, H. (1961). *Theory of Probability,* third edition. Oxford: Clarendon Press.

JUSTEL, A. & PEÑA, D. (1996). Gibbs Sampling will fail in outlier problems with strong masking. *Journal of Computational and Graphical Statistics* (forthcoming).

PEÑA, D. & GUTTMAN, I. (1993). Comparing probabilistic methods for outlier detection in linear models. *Biometrika,* **80,** 603–610.

PEÑA, D. & TIAO, G.C. (1992). Bayesian robustness functions for linear models. In *Bayesian Statistics 4*, Ed. J.M. BERNARDO, J.O. BERGER, A.P. DAWID and A.F.M. SMITH, pp. 365–388. Oxford University Press.

PEÑA, D. & YOHAI, V.J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society B,* **57,** 145–156.

PETTIT, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society B,* **52,** 175–184.

PETTIT, L. (1992). Bayes factor for outliers models using the device of imaginary observations. *Journal of the American Statistical Association,* **87,** 541–545.

PETTIT, L.I. & SMITH, A.F.M. (1985). Outliers and influential observations in linear

models. In *Bayesian Statistics 2,* Ed. J.M. BERNARDO, M.H. DEGROOT, D.V. LINDLEY and A.F.M. SMITH, pp. 473–494. Amsterdam: Elsevier.

ROBERT, C.P. (1994). Convergence assessments for Markov Chain Monte-Carlo methods. Working Paper. Crest, Laboratoire de Statistique, Insee, Paris.

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

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

ROUSSEEUW, P.J. & VAN ZOMEREN, B.C. (1990) Unmaskig multivariate outliers and leverage points. *Journal of the American Statistical Association,* **85,** 633–639.

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. & WASSERMAN, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing,* **1,** 105–117.

WEST, M. (1984). Outlier models and prior distribution in Bayesian linear models. *Journal of the Royal Statistical Society B,* **46,** 431–439.