ELSEVIER

Computational Statistics & Data Analysis 36 (2001) 69–84

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Bayesian unmasking in linear models

Ana Justel[a,*] , Daniel Peña[b]

[a] *Department of Mathematics, Universidad Autónoma de Madrid, Spain*
[b] *Department of Statistics and Econometrics, Universidad Carlos III de Madrid, Spain*

Received 1 April 1999; accepted 1 February 2000

Abstract

We propose a Bayesian procedure for multiple outlier detection in linear models which avoids the masking problem. The posterior probabilities of each data point being an outlier are estimated by using an adaptive learning Gibbs sampling method. The idea is to modify the initial conditions of the Gibbs sampler in order to visit the posterior distribution space in a reasonable number of iterations. To find an appropriate vector of initial values we consider the information extracted from the eigenstructure of the covariance matrix of a vector of latent variables. These variables are introduced in the model to capture the heterogeneity in the data. This procedure also overcomes the false convergence of the Gibbs sampling in problems with strong masking. Our proposal is illustrated with some of the examples most frequently used in the literature. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: 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 Bayesian or classical 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 of the masking effect. The masking problem has received very little attention in the Bayesian literature. Masking occurs when one outlier observation is not detected because of the presence of other outliers. Also,

[*] Corresponding author. Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain. Tel.: +34-1-397-5037; fax: +34-1-397-4889.
E-mail address: ana.justel@uam.es (A. Justel).

PII: S 0167-9473(00)00033-5

one good point can be wrongly disregarded due to the effect of the outliers, and this is called the swamping problem.

In this paper we present a Bayesian procedure to overcome the masking problem in multiple outlier detection for linear models. The posterior probabilities of each observation being an outlier are computed by an adaptive Gibbs sampling procedure in three stages. The first stage runs the Gibbs sampling until the Markov chains reach a stable state. The second stage uses the output from the first stage to compute an outlier free subset by analyzing the covariance structure of the model parameters. The third stage runs again the Gibbs sampling by now the initial conditions are adapted by the information provided by the second stage. The output of this third stage is used to identify the outliers and to compute robust estimates of the parameters in the model.

The paper is organized as follows. In Section 2 the Bayesian linear model is considered and a method is presented to find an outlier free subset. Section 3 develops the new adaptive procedure. Section 4 shows the performance of this algorithm in some examples used as a benchmark in the literature. Some final comments appear in Section 5.

2. Bayesian multiple outlier detection

2.1. Outliers in the Bayesian linear model

Let us consider the Bayesian regression model where the observations $y=(y_1,\dots,y_n)'$ are generated by

$$y_i = x_i'\beta + u_i, \quad i=1,\dots,n, \tag{2.1}$$

where n is the sample size, $X=(x_1,\dots,x_n)'$ is a $n\times m$ matrix of nonrandom variables, β is a $m\times 1$ vector of unknown parameters, and $u=(u_1,\dots,u_n)'$ is a vector of nonobservable perturbations with distribution $\mathrm{N}(0,\sigma^2 I)$. We assume independent and noninformative prior distributions for the location and scale parameters, $P(\beta,\sigma^2)\propto\sigma^{-2}$. 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 most often used diagnostic methods analyse if one observation is compatible with the rest of the sample by studying the predictive distribution $P(y_i|y_{(i)})$, where $y_{(i)}$ is the sample when deleting the data y_i (see Chaloner and Brant (1988), for an alternative approach). The robust methods suppose heavy tail distributions for the errors or mixtures of distributions. The more frequently analysed model is the normal scale contamination model, where the error distribution is

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

The mixture distribution (2.2) indicates that there exists a probability α of each data point being spuriously generated from an alternative distribution. Data points generated from the alternative distribution will be considered as outliers. Assuming that k and α are known, the posterior probability that there are exactly n_I outliers in a set indexed by $I = \{i_1, \dots, i_{n_I}\}$ and $n - n_I$ good points is given by

$$p_I \propto \left(\frac{\alpha}{1-\alpha}\right)^{n_I} k^{-n_I} \left(\frac{|X'X|}{|X'X - \phi X_I' X_I|}\right)^{1/2} \left(\frac{s^2}{s^2_{(I)}}\right)^{(n-p)/2}, \tag{2.3}$$

where $\phi = 1 - k^{-2}$, X_I is the $n_I \times m$ submatrix of X with the rows indexed by I, s^2 is the usual unbiased residual variance estimate and $s^2_{(I)}$ is the estimate computed by considering the n_I points in I generated from the alternative distribution (for the detailed expressions, see Box and Tiao, 1968).

The predictive ordinates $P(y_i \mid y_{(i)})$ and the probabilities (2.3) can be easily used to check for a single outlier in a sample, as well as for checking the presence of a particular group of outliers. However, the most relevant problem is when the number and the position of the outliers are unknown, as it is the usual case with real data. In this case, two ideas may be considered: (1) using the deleting one observation procedure to detect outliers one by one, and (2) to identify multiple outliers by computing all the probabilities for the possible group of outliers. These two possibilities present serious problems in some particular situations. The deleting one by one observation procedure can be subject to masking in samples with multiple outliers, whereas the multiple detection using (2.3) may avoid masking, but they involve the extensive computations of the 2^n posterior probabilities which correspond to all the possible groups. To reduce these computations Peña and Tiao (1992) propose a method based on stratified sampling in the context of building the Bayesian robustness curves BROC and SEBROC. Alternatively, we consider MCMC methods to reduce the computations and to propose an unmasking procedure which can be satisfactorily implemented in moderate and large samples.

2.2. Gibbs sampling for the outlier regression model

Verdinelli and Wasserman (1991) propose to apply the Gibbs sampling to the detection of univariate outliers in a normal random sample and they show that this algorithm overcomes the heavy computations needed in this type of problems. Justel and Peña (1996a) 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 probabilities.

In this paper we generalize the normal scale contamination model (2.1) and (2.2) by assuming for the contamination parameter α a prior distribution Beta(γ_1, γ_2) with expectation $\alpha_0 = E(\alpha) = \gamma_1/(\gamma_1 + \gamma_2)$. The application of the Gibbs sampling (see Gilks et al., 1996) is carried out by augmenting the parameter vector with a set of classification variables $\delta = (\delta_1, \dots, \delta_n)'$, defined as $\delta_i = 1$ when y_i is generated by the alternative distribution $\mathrm{N}(x_i'\beta, k^2\sigma^2)$, and $\delta_i = 0$ otherwise. The pair (y_i, x_i') will be called an outlier when the marginal probability $p_i = P(\delta_i = 1 \mid y)$ is greater than 0.5. Thus, α is the prior probability that any observation is an outlier. The full conditional distributions are: (1) the conditional distribution of β is $\mathrm{N}_m(\tilde{\beta}, \sigma^2(X'V^{-1}X)^{-1})$, where

$\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$ and V is a diagonal matrix with elements $v_{ii} = 1 + \delta_i(k^2 - 1)$; (2) the conditional distribution of σ^2 is *Inverted-Gamma* $(n/2, \sum u_i^{*^2}/2)$, where $u_i^* = (y_i - x_i'\beta)/(1 + \delta_i(k-1))$; (3) the conditional distribution of α only depends on the vector δ and is *Beta* $(\gamma_1 + \sum \delta_i, \gamma_2 + n - \sum \delta_i)$; and (4) the conditional distribution of δ_i is *Bernoulli* with success probability

$$P(\delta_i = 1 \mid y, \beta, \sigma^2, \alpha) = \left(1 + \left(\frac{1-\alpha}{\alpha}\right) F_{10}(i)\right)^{-1}, \tag{2.4}$$

where F_{10} is the Bayes factor given by

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

and $\phi = 1 - k^{-2}$. Note that the conditional probability that the ith observation is an outlier depends only on u_i^2/σ^2. If u_i is small, $F_{10}(i)$ will be large and the probability (2.4) will be small. The opposite occurs when u_i^2/σ^2 is large.

Justel and Peña (1996a) showed in several examples that Gibbs sampling will fail for outlier detection in data sets with masking problems. The lack of convergence in these cases seems to be due to the effect of the leverage. 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 $S_0 = (y_0, X_0)$ be the set of observations classified as good in the initial conditions. From now on, the subscript (I) means that the data indexed by I are deleted. Then we assume the following *initial condition dependence property*:

(a) 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. To justify this, consider the case in which S_0 includes the group of outliers. It can be proved (see Peña and Yohai, 1995) that the computed error for the outliers, $u_i^{(0)}$, can be expressed as

$$u_i^{(0)} = \frac{y_i - x_i'\beta_{(I)}^{(0)}}{1 + n_I h_i} \qquad \text{for } i \in I, \tag{2.5}$$

where $h_i = x_i'(X_{0(I)}'X_{0(I)})^{-1}x_i$ is the leverage of the outliers and $\beta_{(I)}^{(0)}$ is the mean of the conditional distribution given $\delta^{(0)}$, both computed when the data indexed by I are deleted. For large k, $\beta_{(I)}^{(0)}$ is approximately the least-squares estimate with the subsample S_0 where the observations indexed by I are deleted. As h_i is unbounded, $u_i^{(0)}$ will be small if h_i is large and this effect increases with the number of outliers n_I. Therefore, for high leverage outliers $u_i^{(0)}$ will be close to zero and so will be the probability (2.4).

(b) if S_0 includes no outliers, the existing outliers are always identified, and the good data are not misspecified. This will happen because if the set S_0 does not contain outliers, $u_i^{(0)} = y_i - x_i'\beta^{(0)}$ will be large for $i \in I$, and the probability (2.4) will be close to one.

Therefore, a clear objective is to find 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; Hawkins et al., 1984).

2.3. Finding an outlier free subset

Two outliers are masked when they need to be identified as such jointly. Suppose that the sample includes two or more very strongly masked outliers. This means that the probability of identifying the ith observation as an outlier subject to the condition that the jth observation is an outlier must be close to one, because either all the outliers are identified jointly or none of them is detected as an outlier. Hence, $P(\delta_i=1 \mid \delta_j=1, y) \simeq 1$, where δ_i, δ_j are the classification variables corresponding to these two outliers. Let us call $p=P(\delta_j=1 \mid y)$, and let us assume that, approximately, also $P(\delta_i=1 \mid y)=p$. This implies that $P(\delta_i=1, \delta_j=1 \mid y) \simeq P(\delta_j=1 \mid y)=p$, and the covariance between the binary variables δ_i and δ_j is

$$c_{ij}=P(\delta_i=1, \delta_j=1 \mid y)-P(\delta_i=1 \mid y)P(\delta_j=1 \mid y)=p-p^2$$

and if p is small this covariance will be of order p. Suppose now that δ_i, δ_j correspond to two good points. Then we expect that $P(\delta_i=1 \mid \delta_j=1, y) \simeq P(\delta_i=1 \mid y)$ and, therefore, c_{ij} will be close to zero. Finally, if δ_i is an outlier and δ_j a good observation, as $P(\delta_i=1 \mid \delta_j=1, y) \simeq P(\delta_i=1 \mid y)$, again the covariance c_{ij} will be close to zero.

Consider now the estimation with the Gibbs sampling. Suppose that we select the initial conditions S_0 in such a way that the probability that S_0 is outlier free is q (we will discuss how to do it in Section 3). Then we run R parallel sequences and let us call $\hat{\delta}^r=(\hat{\delta}_1^r, \ldots, \hat{\delta}_n^r)'$, $r=1, \ldots, R$, to the vectors of last generated values for the classification variables, which are used to estimate the marginal outlier posterior probabilities by

$$\hat{p}_i=\frac{1}{R}\sum_{r=1}^{R}\hat{\delta}_i^r,$$

and the joint probabilities by

$$\hat{p}_{ij}=\frac{1}{R}\sum_{r=1}^{R}\hat{\delta}_i^r\hat{\delta}_j^r.$$

In the general case with n_I masked outliers, the sample covariance between two classification variables, $\hat{\delta}_i=(\hat{\delta}_i^1, \ldots, \hat{\delta}_i^R)'$ and $\hat{\delta}_j=(\hat{\delta}_j^1, \ldots, \hat{\delta}_j^R)'$, will also be large. This is so because $\hat{\delta}_i^r$ and $\hat{\delta}_j^r$ will only be 1 when all the other classification variables for the rest of the outliers are also 1, and both will be zero otherwise. Then the sample covariance $\hat{c}_{ij}=\hat{p}_{ij}-\hat{p}_i\hat{p}_j$ will be higher for masked outliers than for good data points or combinations of one good point and an outlier.

Let $D=(\hat{\delta}^1, \ldots, \hat{\delta}^R)'$ be the data matrix for the classification variables. The covariance matrix $\hat{C}$ is

$$\hat{C}=\frac{1}{R}D'M_RD,$$

where $M_R=I-1_R1_R'/R$, and $1_R'=(1, \ldots, 1)$. Let us consider the expected behaviour of the eigenvectors and eigenvalues of $\hat{C}$ in the limit case in which the first $n_G>0$ observations correspond to good data, the next $n_H \geq 0$ to good data that are swamped and the last $n_I > 0$ to the set of outliers. Also let us assume that the first Q runs correspond to the runs in which the set S_0 is outlier free and therefore the outliers are correctly identified. Then $\hat{\delta}^{r} = (0', 0', 1'_I)'$ for $r = 1, \ldots, Q$, where the first n_G correspond to the good data, the next n_H to the swamped data, and the final n_I to the outliers that are correctly identified. Also, $\hat{\delta}^{r} = (g'_r, 1'_H, 0')'$ for $r = Q+1, \ldots, R$, where the swamped good data are identified as outliers and the vector g_r may contains a few nonnull elements because the outlier probability for good data is small, but not zero. We may suppose that there are not important differences between these columns in the proportion of ones (misspecifications), and that this number is bounded by some small value π, such that

$$\frac{1}{R}\sum_{i=1}^{R-Q} g_{ij} \leq \pi \quad \text{for all } j = 1, \ldots, n_G. \tag{2.6}$$

In summary, the matrix D will be

$$D = \begin{pmatrix} 0 & \vdots & 0 & \vdots & 1_Q 1'_{n_I} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ G & \vdots & 1_{\bar{Q}} 1'_{n_H} & \vdots & 0 \end{pmatrix},$$

where $\bar{Q} = R - Q$ and $G = (g_{Q+1}, \ldots, g_R)'$ is a matrix $\bar{Q} \times n_G$. The covariance matrix $\hat{C}$ can be written as

$$\hat{C} = \begin{pmatrix} \dfrac{1}{R} G' M_{\bar{Q}} G & \vdots & \dfrac{Q}{R^2} G' 1_{\bar{Q}} 1'_{n_H} & -\dfrac{Q}{R^2} G' 1_{\bar{Q}} 1'_{n_I} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{Q}{R^2} 1_{n_H} 1'_{\bar{Q}} G & \vdots & \dfrac{Q\bar{Q}}{R^2} 1_{n_H} 1'_{n_H} & -\dfrac{Q\bar{Q}}{R^2} 1_{n_H} 1'_{n_I} \\ -\dfrac{Q}{R^2} 1_{n_I} 1'_{\bar{Q}} G & \vdots & -\dfrac{Q\bar{Q}}{R^2} 1_{n_I} 1'_{n_H} & \dfrac{Q\bar{Q}}{R^2} 1_{n_I} 1'_{n_I} \end{pmatrix}.$$

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

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

where $\hat{C}_{22}$ is the $(n_H + n_I) \times (n_H + n_I)$ matrix

$$\hat{C}_{22} = \frac{Q\bar{Q}}{R^2} \begin{pmatrix} 1_{n_H} 1'_{n_H} & \vdots & -1_{n_H} 1'_{n_I} \\ \cdots & \cdots & \cdots \\ -1_{n_I} 1'_{n_H} & \vdots & 1_{n_I} 1'_{n_I} \end{pmatrix}.$$

The eigenvalues of $\hat{C}$ are the eigenvalues of the matrices $G'G/R$ and $\hat{C}_{22}$. By Eq. (2.6) the eigenvalues of $G'G/R$ satisfy

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

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

$$\lambda_I = q(1-q)(n_H + n_I). \tag{2.7}$$

Then the matrix $\hat{C}$ 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, $v_a = (0'_{n_G}, a1'_{n_H}, -a1'_{n_I})'$ is an eigenvector of the matrix $\hat{C}$ associated with λ_I, for all nonnull values of a.

The λ_I eigenvalue, given by (2.7) 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 q close to zero corresponds to the strong contamination case. A value q close to one corresponds to the case in which there are no outliers in the sample or only very few. In this case, the outliers will not be masked and they can be directly detected by the Gibbs sampling. The interesting case is when $0 < q < 1$ and n_I (and maybe n_H) is large. This corresponds to the most difficult case in which outliers are not identified in most runs. 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 coefficients (in absolute value) on the eigenvector v_a are potentially outlier candidates. As a result, we may split the data into two subsets: (1) the set that contains the observations with nonnull coefficients on the eigenvector v_a or with a high individual probability $\hat{p}_i$; and (2) the set of the remaining observations. We call the first set the *potential outlier set* (PO).

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

2.4. Example

The artificial data set proposed by Hawkins et al. (1984) is a well-known example of masking. We will call this dataset the HBK data and it is represented by a matrix plot in Fig. 1. Out of the 75 observations in four dimensions, data from 1 to 10 are high-leverage outliers. The traditional outlier identification procedures based on least-squares estimation are not able to identify these outliers due to their high leverage. In addition, observations 11–14 are good data wrongly identified as outliers. Justel and Peña (1996a) show that Gibbs sampling fails with this data set. The 10 outliers are not identified and the Gibbs sampling suffers the same problems as traditional methods for outlier detection.

The covariance matrix for the HBK data is shown in Table 1. As expected, the covariance is large and positive for the masked outliers and the swamped good data. The covariance is large and negative between one masked outlier and one swamped good data. The largest eigenvalues of this matrix are $\lambda_1 = 3.4297$ and $\lambda_2 = 0.0391$. The large difference between these values is corroborated by the percentages of variance explained by the eigenvalues, these are 78.5% for λ_1 and 0.9% for λ_2. The

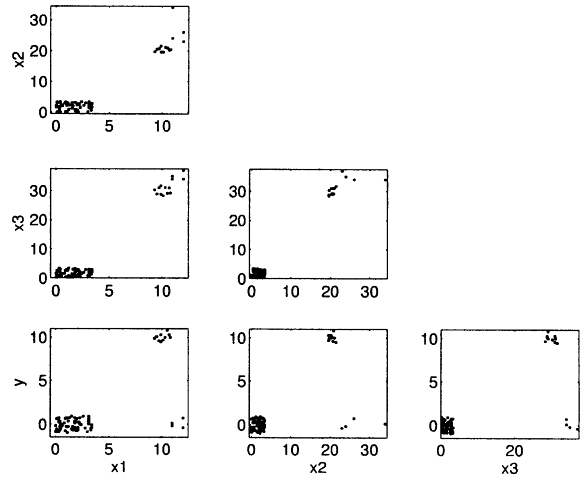

Fig. 1. Matrix plot of the HBK data.

Table 1
Covariance matrix with HBK data. Values greater than 0.01 and less than −0.01 are printed

2	0.22																
3	0.23	0.22															
4	0.22	0.22	0.22														
5	0.22	0.22	0.23	0.22													
6	0.23	0.22	0.23	0.22	0.23												
7	0.22	0.22	0.22	0.21	0.22	0.22											
8	0.22	0.22	0.22	0.22	0.22	0.22	0.21										
9	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22									
10	0.23	0.22	0.23	0.22	0.23	0.23	0.22	0.22	0.22								
11	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22							
12	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	0.22						
13	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.21	−0.22	−0.22	−0.22	0.22	0.22					
14	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	−0.22	0.22	0.22	0.22				
15	.	.	.	.	.	.	.	.	.	.	.	.	.				
16	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

components of the eigenvector associated with the highest eigenvalue are shown in Fig. 2. As a result, we shall include in PO the observations 1–14.

The matrix $\hat{C}$ was built with the estimated probabilities after 500 iterations of $R = 300$ parallel sequences of the Gibbs sampling. Each sequence started with a set S_0 of four observations considered as good data points. Note that here $n_I = 10$, $n_H = 4$ and the probability of no outliers in S_0 is

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

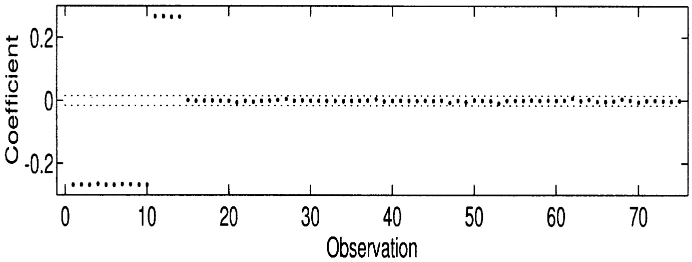

Fig. 2. Coefficients of the eigenvector associated with the eigenvalue λ_1 of the covariance matrix with HBK data.

Therefore, the expected value of the largest eigenvalue is, according to (2.7) equal to 3.45, that is very similar to the real observed value.

3. Posterior probabilities sampling algorithm

We propose an Adaptive Gibbs Sampling Algorithm in which the initial conditions of the Gibbs sampler are changed according to a three-stage procedure. In the first stage, the Gibbs sampling is initialized by using a small set of observations classified to be good. Then the algorithm is run for a few iterations. In the second stage, the dependency among the classification variables computed from the run is taken into account in order to find the potential outlier set (PO). In the third stage the Gibbs sampler is initialized giving value 1 to the classification variables in PO, and it is again run for a few iterations. Using the values in the last iterations of only one sequence, inference from this sample allows us to identify the outliers and to estimate the parameters in the model. Accordingly, we suggest to follow these stages:

Stage 1 (*Standard Gibbs sampler*): Run the Gibbs sampling in parallel until the Markov chains reach a stable state. The initial conditions for each sequence are selected as follows:

(i) Let n_0 be the maximum integer such that the probability of finding at most one outlier in any data subset of size n_0 is greater than c_1. Then select randomly ℓ data points $y_{i_1},\dots,y_{i_\ell}$, where $\ell=\max\{n_0,m\}$. The initial set is

$$S_0=\{(y_{i_1},x'_{i_1}),\dots,(y_{i_\ell},x'_{i_\ell})\}\setminus\{\text{single outliers detected by diagnostic tools}\}.$$

(ii) The initial conditions are:

(a) $\delta_j^{(0)}=1$ for all $(y_j,x'_j)\in S_0$, and $\delta_j^{(0)}=0$ otherwise.

(b) $\beta^{(0)}=(X'V^{(0)^{-1}}X)^{-1}X'V^{(0)^{-1}}y$, where $V^{(0)}$ is a diagonal matrix with $v_{jj}^{(0)}=1+\delta_j^{(0)}(k^2-1)$.

The initial conditions of Stage 1 are such that with high probability the initial set S_0 is outlier free. The decision about the size of S_0 is 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 a set of size m. To compute n_0 we must consider that 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. Let q_1 be the probability that the set S_0 contains at most one outlier, then

$$q_1 = \binom{\bar{n}_\alpha}{n_0}\binom{n}{n_0}^{-1} + \binom{\bar{n}_\alpha}{n_0-1}\binom{n_\alpha}{1}\binom{n}{n_0}^{-1},$$

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 prior expectation of the parameter α, and given α_0 and n_0 we can compute q_1. The value n_0 is chosen so that $q_1 > c_1$. Finally, the single outliers 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, Appendix B) scale of evidence.

Stage 2 (*Outlier identification*): Estimate the probabilities $\hat{p}_j$ and the covariance matrix $\hat{C}$ with the values of the classification parameters from the last iteration of each sequence. Compute the largest c_2 eigenvalues and associated eigenvectors $(v_1, v_2, \ldots)$. Then the potential outlier set PO contains the data (y_j, x_j') such that $\hat{p}_j > 0.5$, or $|v_{ij}| > c_3\, m_i$, for any $i = 1, \ldots, c_2$, where $m_i = \text{median}_j\{|v_{ij}|\}/0.6475$. (The value 0.6475 is chosen in agreement with standard practice in robust statistics in order to make the robust scale estimate consistent for the normal distribution, see Huber, 1981.)

Stage 3 (*Estimation*): Reset the algorithm and run the Gibbs sampling once, until the Markov chain reaches a stable state. The initial conditions are:

(a) $\delta_j^{(0)} = 1$ for all $(y_j, x_j') \in$ PO, and $\delta_j^{(0)} = 0$ otherwise.

(b) $\beta^{(0)} = (X'V^{(0)^{-1}}X)^{-1}X'V^{(0)^{-1}}y$, where $V^{(0)}$ is a diagonal matrix with $v_{jj}^{(0)} = 1 + \delta_j^{(0)}(k^2-1)$.

The output of the Gibbs sampling is used to estimate the posterior probabilities of all parameters in the model.

The bounds c_1 and c_2 and the constant c_3 must be chosen. As c_1 is a bound for the probability that the initial set is outlier-free we suggest values around 0.95 in order to consider both sensitivity and power. For c_2 we choose the minimum value of (m, c_2^*), where c_2^* is the number of eigenvalues greater than five times a robust dispersion measure of the eigenvalues λ_i of $\hat{C}$, that can be $median\{\lambda_i\}/0.6475$. For the constant c_3 we have chosen the value 3, so that we consider interesting points those that are larger than three standard deviations. We have checked that small changes of these two last constants do not affect the results of the algorithm. The number of iterations needed to reach the stabilization in both stages may be decided by the methods for monitoring convergence proposed by Gelman and Rubin (1992) or Robert (1995, 1998), 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 $\varepsilon > 0$, $|\hat{p}_i^{(S-1)} - \hat{p}_i^{(S)}| < \varepsilon$ for all $i = 1, \ldots, n$. Finally, in Stage 2 the initial conditions are always the same and we run only one sequence to reduce the computational effort.

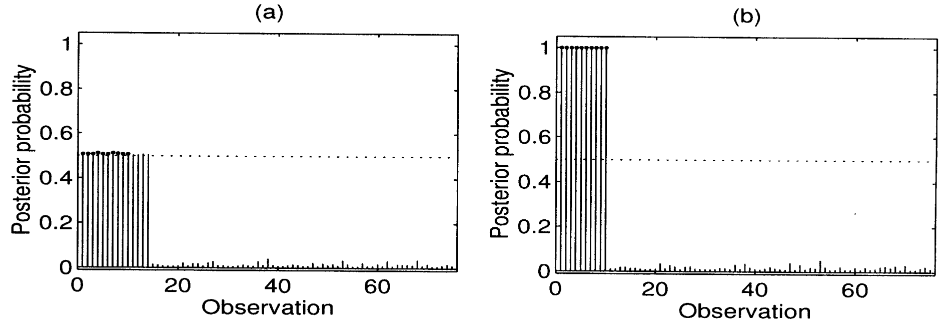

Fig. 3. Results of the Gibbs sampler with HBK data: (a) probabilities of each data point to be outlier in the Stage 1; (b) posterior outlier probabilities in the Stage 3.

4. Procedure performance

The Adaptive Gibbs Sampling Algorithm (AGSA) has been tested with many real datasets presented in the literature of outliers and all cases it has led to the expected good solution (some of these examples are presented in Justel and Peña, 1996b). In order to show that it can also work on very difficult data sets we have considered two well known simulated examples that are often used as benchmark to judge the power of outlier detection procedures. These examples are the HBK data, presented in Section 2.4 and the Rousseeuw (1984) data presented in the introduction. In the two examples, a Gibbs sampler with 300 sequences is used and the number of iterations is decided with $\varepsilon = 0.002$. In all the examples $\alpha_0 = 0.2$ and $\gamma_1 + \gamma_2 = n$, that imply $E(\alpha \mid \delta) = \frac{1}{2}E(\alpha) + \frac{1}{2}\bar{\delta}$, and $k = 10$.

4.1. HBK data

The procedure is applied to the HBK data discussed in Section 2.4. The observations 1–10 are outliers which swamp the good data 11–14.

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 shown in Fig. 2. In this example the estimates of the individual probabilities, shown in Fig. 3(a), and the eigenstructure of the covariance matrix, discussed in Section 2.4, lead to the same conclusion: the group of potential outliers PO includes the observations 1–14, that are the masked outliers and the swamped good data. In Stage 3, these data points are considered outliers in the initial conditions and the outliers are correctly identified with probability equal to one (see Fig. 3(b) for the posterior outlier probabilities). Note that the probabilities are very low for the four previously swamped data.

4.2. Rousseeuw-type data

The set of simulated data proposed by Rousseeuw (1984) is generated in two groups, that can be seen in the scatter plot of Fig. 4. See Table 2 for the numerical

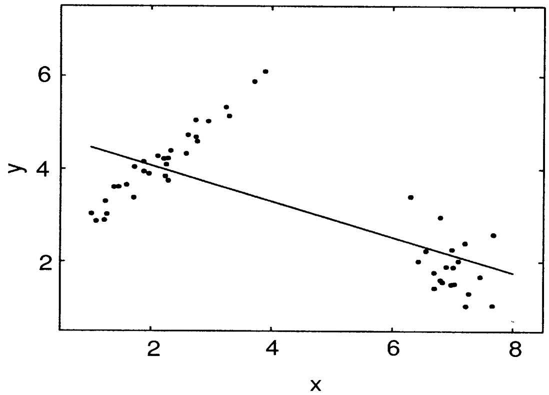

Fig. 4. Rousseeuw-type data.

Table 2
Rousseeuw type data

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

results. One group follows the linear model y_i=2+x_i+u_i with error standard deviation 0.2, whereas the other group comes from a bivariate normal with mean (7,2) and covariance matrix 0.5I. Then out of the 50 data points, 20 are high-leverage outliers (data 1–20) and 30 are good observations (data 21–50).

This is a difficult example since the contamination is 40%, and many times it is used as a benchmark for the robust estimation methods and the diagnostic procedures for outlier detection. The usual diagnostic procedures identify as outliers the observations 32 and 33, which are good data with large least squares residuals. The solid line in the Fig. 4 is the least-squares estimator of the regression line. Also the standard Gibbs sampler does not identify the outliers, as shown by Justel and Peña (1996a).

The AGSA proposed in this paper works very well. Starting with a set of four good observations, the outlier probabilities in Stage 1 for the 20 outliers are low (see Fig. 5a), but the covariance matrix has two nonnull eigenvalues $\lambda_1 = 0.53$ and $\lambda_2 = 0.31$. The coordinates of the associated eigenvectors are shown in Fig. 6. In the first eigenvector the results are as expected: (1) the coordinates are nonnull for

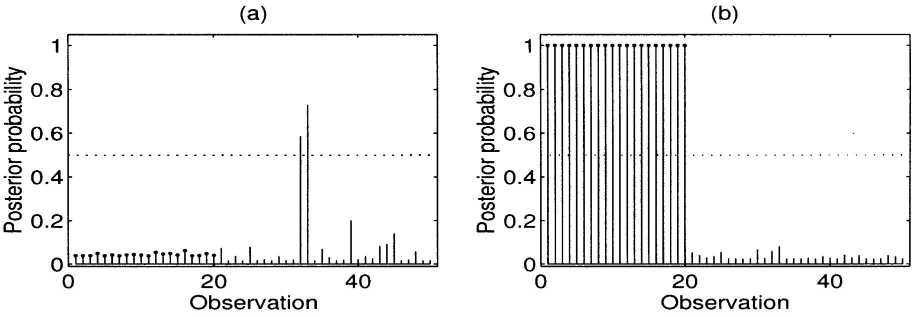

Fig. 5. Results of the Gibbs sampler with Rousseeuw-type data: (a) probabilities of each data point to be outlier in the Stage 1; (b) posterior outlier probabilities in the Stage 3.

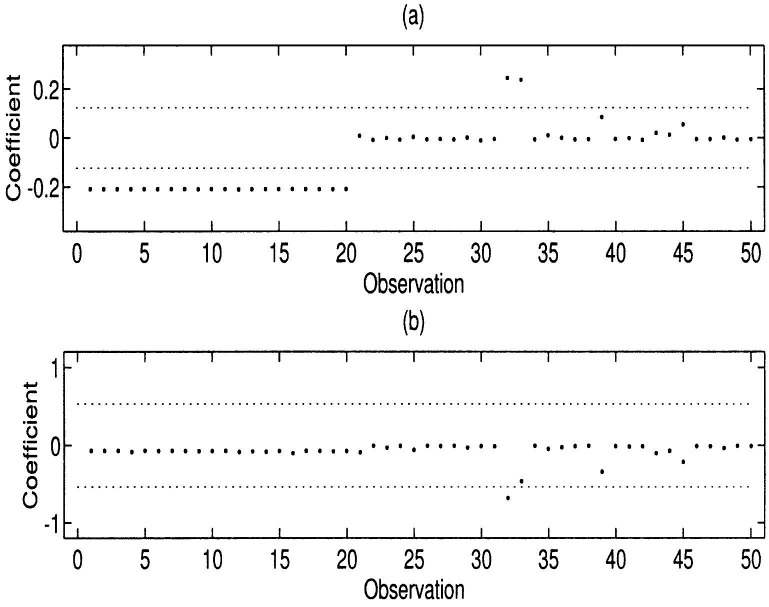

Fig. 6. Coefficients of the eigenvectors associated with the eigenvalues λ_1 (in (a)) and λ_2 (in (b)) of the covariance matrix with Rousseeuw-type data.

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 observations 32 and 33. The posterior outlier probabilities estimated in the second stage (see Fig. 5b) are such that the outliers are correctly identified in a few iterations and also the swamping effect disappears.

In this example, the outliers could be identified because they have larger variability compared to the good data. Let us analyse what happens if the cluster of 20 "bad" data is generated from a bivariate normal with the same mean as before but now the standard deviation is $0.2I$. The data in Table 3 have maintained the same variability

Table 3
Rousseeuw-type modified data

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
x	6.89	7.17	6.95	6.85	6.91	6.87	6.95	7.11	6.78	7.07
y	2.07	2.12	2.17	2.35	1.93	1.97	2.12	2.19	1.77	1.88
x	7.22	7.02	7.14	6.98	7.40	7.18	6.63	7.00	6.63	7.20
y	2.00	1.50	2.23	1.79	2.23	1.84	2.12	2.16	1.96	2.11

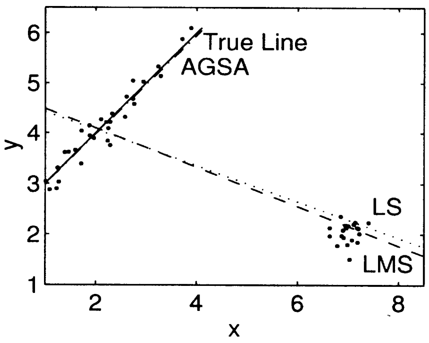

Fig. 7. Rousseeuw modified data, true regression line, LS estimate, LMS estimate, AGSA estimate.

for the good points and the cluster of outliers. Fig. 7 shows the data, the true line $y=2+x$, the least-squares line (LS), the least median of square line (LMS) computed from 10^5 replications, and the line obtained by the AGSA.

Fig. 7 shows that the procedure proposed in this paper is not affected by this change. We observe that the outlier probabilities in the first stage do not allow us to identify the outliers and the nonnull eigenvalues of the covariance matrix are now $\lambda_1 = 2.2$ and $\lambda_2 = 0.77$. In the first eigenvalues the results are as before and again the outliers are identified in the third stage and we obtain a robust regression estimation. Note that the bad performance of LMS does not depend on the number of replications because the median of the squared residuals is 0.0992 for the true line, 0.0959 for our procedure, and 0.0622 for the LMS regression line.

5. Concluding remarks

The Bayesian procedure for outlier detection in linear models presented in this paper combines a sequential learning procedure for Gibbs sampling with the information from an estimate of the covariance matrix of the classification variables. The eigenvectors associated to the nonzero eigenvalues of this matrix provide information about which observations 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 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.

We have assumed k as known but the procedure can be easily extended by: (a) assuming k unknown and introducing a prior distribution over its possible values, or (b) considering $\sigma_1 = \sigma$ and $\sigma_2 = k\sigma$, with $\sigma_1 < \sigma_2$ (to avoid *label-switching*). In this case we should assume proper prior distributions, as the conjugate priors, to avoid improper posterior distributions. Other simple possibility is to do a sensitivity analysis changing the value of k. We have found that in real datasets the results are fairly robust to a sensible value of k in the range $(5, 15)$.

There are other procedures which use the eigenstructure of some matrix in order to identify outliers avoiding the masking problem (see, for instance, Jorgensen, 1992; Peña and Yohai, 1995). Our procedure has the advantage of using the covariance matrix of the identifying variables, which has a clear justification as indicated in Section 2.2.

Acknowledgements

We are grateful to Luc Bauwens and two referees for very useful comments. Part of the manuscript was prepared when the first author was Research Associate at CORE, Université Catholique de Louvain. She was financed by the European Commission in the TMR Programme. Second author is partially financed by Cátedra BBV de Calidad. Both authors would like to acknowledge support for this research provided by DGES (Spain), under Grants PB97-0021 (A. Justel) and PB96-0111 (D. Peña).

References

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

Chaloner, K., Brant, R., 1988. A Bayesian approach to outlier detection and residual analysis. Biometrika 75, 651–659.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman & Hall, New York.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences (with discussion). Statist. Sci. 7, 457–511.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

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

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Jeffreys, H., 1961. Theory of Probability, 3rd Edition. Clarendon Press, Oxford.

Jorgensen, B., 1992. Finding rank leverage subsets in regression. Scand. J. Statist. 19, 139–156.

Justel, A., Peña, D., 1996a. Gibbs Sampling will fail in outlier problems with strong masking. J. Comput. Graphical Statist. 5, 176–189.

Justel, A., Peña, D., 1996b. Bayesian unmasking in linear models. CORE Discussion Paper 9619, Université Catholique de Louvain.

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: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4. Oxford University Press, Oxford, pp. 365–388.

Peña, D., Yohai, V.J., 1995. The detection of influential subsets in linear regression by using an influence matrix. J. Roy. Statist. Soc. B 57, 145–156.

Pettit, L.I., Smith, A.F.M., 1985. Outliers and influential observations in linear models. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), Bayesian Statistics 2. Elsevier, Amsterdam, pp. 473–494.

Robert, C.P., 1995. Convergence control methods for Markov Chain Monte Carlo algorithms. Statist. Sci. 10, 231–253.

Robert, C.P. (Ed.), 1998. Discretization and MCMC Convergence Assessment. Lecture Notes in Statistics, vol. 135. Springer, New York.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Amer. Statist. Assoc. 79, 871–880.

Verdinelli, I., Wasserman, L., 1991. Bayesian analysis of outlier problems using the Gibbs sampler. Statist. Comput. 1, 105–117.