Statistica Sinica **11**(2001), 651-673

DETECTION OF OUTLIER PATCHES IN AUTOREGRESSIVE TIME SERIES

Ana Justel, Daniel Peña and Ruey S. Tsay

Universidad Autónoma de Madrid, Universidad Carlos III de Madrid and University of Chicago

Abstract: This paper proposes a procedure to detect patches of outliers in an autoregressive process. The procedure is an improvement over the existing detection methods via Gibbs sampling. We show that the standard outlier detection via Gibbs sampling may be extremely inefficient in the presence of severe masking and swamping effects. The new procedure identifies the beginning and end of possible outlier patches using the existing Gibbs sampling, then carries out an adaptive procedure with block interpolation to handle patches of outliers. Empirical and simulated examples show that the proposed procedure is effective.

Key words and phrases: Gibbs sampler, multiple outliers, sequential learning, time series.

1. Introduction

Outliers in a time series can have adverse effects on model identification and parameter estimation. Fox (1972) defined additive and innovative outliers in a univariate time series. Let $\{x_t\}$ be an autoregressive process of order p, AR(p), satisfying

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t, \tag{1.1}$$

where $\{a_t\}$ is a sequence of independent and identically distributed Gaussian variables with mean zero and variance σ_a^2, and the polynomial $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ has no zeros inside the unit circle. An observed time series y_t has an additive outlier (AO) at time T of size β if it satisfies $y_t = \beta I_t^{(T)} + x_t$, $t = 1, \ldots, n$, where $I_t^{(T)}$ is an indicator variable such that $I_t^{(T)} = 0$ if $t \neq T$, and $I_t^{(T)} = 1$ if $t = T$. The series has an innovative outlier (IO) at time T if the outlier directly affects the noise process, that is, $y_t = \phi(B)^{-1}\beta I_t^{(T)} + x_t$, $t = 1, \ldots, n$. Chang and Tiao (1983) show that additive outliers can cause serious bias in parameter estimation whereas innovative outliers only have minor effects in estimation. We deal with additive outliers that occur in patches in an AR process. The main motivation for our study is that multiple outliers, especially those which occur

closely in time, often have severe masking effects that can render the usual outlier detection methods ineffective.

Several procedures are available in the literature to handle outliers in a time series. Chang and Tiao (1983), Chang, Tiao and Chen (1988) and Tsay (1986, 1988) proposed an iterative procedure to detect four types of disturbance in an autoregressive integrated moving-average (ARIMA) model. However, these procedures may fail to detect multiple outliers due to masking effects. They can also misspecify "good" data points as outliers, resulting in what is commonly referred to as the swamping or smearing effect. Chen and Liu (1993) proposed a modified iterative procedure to reduce masking effects by jointly estimating the model parameters and the magnitudes of outlier effects. This procedure may also fail since it starts with parameter estimation that assumes no outliers in the data, see Sánchez and Peña (1997). Peña (1987, 1990) proposed diagnostic statistics to measure the influence of an observation. Similar to the case of independent data, influence measures based on data deletion (or equivalently, using techniques of missing value in time series analysis) will encounter difficulties due to masking effects.

A special case of multiple outliers is a patch of additive outliers. This type of outliers can appear in a time series for various reasons. First and perhaps most importantly, as shown by Tsay, Peña and Pankratz (1998), a multivariate innovative outlier in a vector time series can introduce a patch of additive outliers in univariate marginal times series. Second, an unusual shock may temporarily affect the mean and variance of a univariate time series in a manner that cannot be adequately described by the four types of outlier commonly used in the literature, or by conditional heteroscedastic models. The effect is a patch of additive outliers. Because outliers within a patch tend to interact with each other, introducing masking or smearing, is important in applications to detect them. Bruce and Martin (1989) were the first to analyze patches of outliers in a time series. They proposed a procedure to identify outlying patches by deleting blocks of consecutive observations. However efficient procedures to determine the block sizes and to carry out the necessary computation have not been developed.

McCulloch and Tsay (1994) and Barnett, Kohn and Sheather (1996, 1997) used Markov Chain Monte Carlo (MCMC) methods to detect outliers and compute the posterior distribution of the parameters in an ARIMA model. In particular, McCulloch and Tsay (1994) showed that the Gibbs sampling provides accurate parameter estimation and effective outlier detection for an AR process when the additive outliers are not in patches. However, as clearly shown by the following example, the usual Gibbs sampling may be very ineffcient when the outliers occur in a patch.

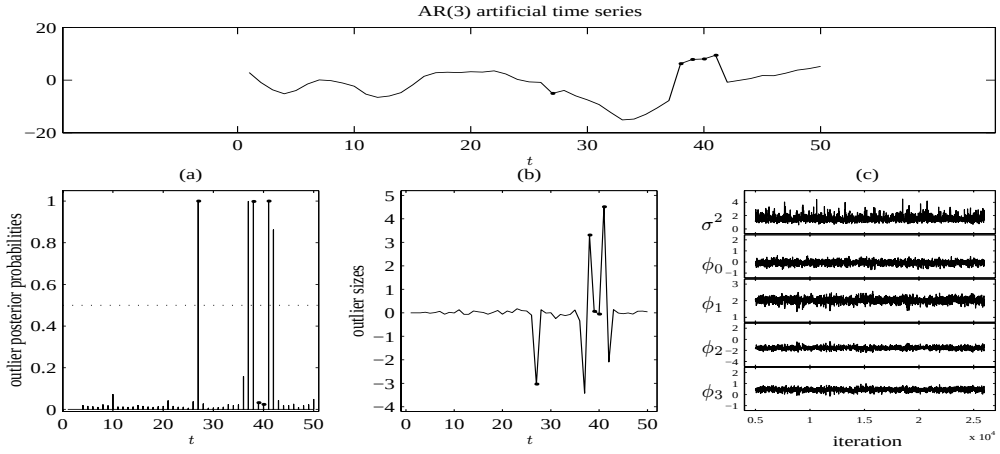


Figure 1. *Top:* AR(3) artificial time series with five outliers at periods $t =$ 27 and 38–41 (marked with a dot). *Bottom:* results of the Gibbs sampler after 26,000 iterations; (a) posterior probabilities for each data point to be outlier; (b) posterior mean estimates of the outlier sizes for each data; and (c) convergence monitoring by plotting the parameter values drawn in each iteration.

Consider the outlier-contaminated time series shown in Figure 1. The outlier-free data consist of a random realization of $n = 50$ observations generated from the AR(3) model,

$$x_t = 2.1x_{t-1} - 1.46x_{t-2} + 0.336x_{t-3} + a_t \qquad t = 1, \ldots, 50, \tag{1.2}$$

where $\sigma_a^2 = 1$. The roots of the autoregressive polynomial are 0.6, 0.7 and 0.8, so that the series is stationary. A single additive outlier of size -3 has been added to the time index $t = 27$, and a patch of four consecutive additive outliers have been introduced from $t = 38$ to $t = 41$, with sizes $(11, 10, 9, 10)$. The data are available to author request. Assuming that the AR order $p = 3$ is known, we performed the usual Gibbs sampling to estimate model parameters and to detect outliers. Figure 1 gives some summary statistics of the Gibbs sampling output using the last 1,000 samples from a Gibbs sampler of 26,000 iterations (when the Markov chains are stabilized as shown in Figure 1-c). Figure 1-a provides the posterior probabilities of being an outlier for each data point, and Figure 1-b gives the posterior means of outlier sizes. From the plots, it is clear that the usual Gibbs sampler easily detects the isolated outlier at $t = 27$, with posterior probability close to one and posterior mean of outlier size -3.03. Meanwhile, the usual Gibbs sampler encounters several difficulties. First, it fails to detect the inner points of the outlying patch as outliers (the outlying posterior probabilities

are very low at $t = 39$ and 40). This phenomenon is referred to as masking. Second, the sampler misspecifies the "good" data points at $t = 37$ and 42 as outliers because the outlying posterior probabilities of these two points are close to unity. The posterior means of the sizes of these two erroneous outliers are -3.42 and -2.09, respectively. In short, two "good" data points at $t = 37$ and 42 are swamped by the patch of outliers. Third, the sampler correctly identifies the boundary points of the outlier patch at $t = 38$ and 41 as outliers, but substantially underestimates their sizes (the posterior means of the outlier sizes are only 3.31 and 4.51, respectively).

The example clearly illustrates the masking and swamping problems encountered by the usual Gibbs sampler when additive outliers exist in a patch. The objective of this paper is to propose a new procedure to overcome these difficulties. Limited experience shows that the proposed approach is effective.

The paper is organized as follows. Section 2 reviews the application of the standard Gibbs sampler to outlier identification in an AR time series. Section 3 proposes a new adaptive Gibbs algorithm to detect outlier patches. The conditional posterior distributions of blocks of observations are obtained and used to expedite the convergence of Gibbs sampling. Section 4 illustrates the performance of the proposed procedure in two examples.

2. Outlier Detection in an AR Process

2.1. AR model with additive outliers

Suppose the observed data, $\boldsymbol{y} = (y_1, \ldots, y_n)'$ are generated by $y_t = \delta_t\beta_t + x_t$, $t = 1, \ldots, n$, where x_t is given by (1.1) and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$ is a binary random vector of outlier indicators; that is, $\delta_t = 1$ if the tth observation is contaminated by an addtive outlier of size β_t, and $\delta_t = 0$ otherwise. For simplicity, assume that $x_1, \ldots, x_p$ are fixed and $x_t = y_t$ for $t = 1, \ldots, p$, i.e., there exist no outliers in the first p observations. The indicator vector of outliers then becomes $\boldsymbol{\delta} = (\delta_{p+1}, \ldots, \delta_n)'$ and the size vector is $\boldsymbol{\beta} = (\beta_{p+1}, \ldots, \beta_n)'$. Let $\boldsymbol{X}_t = (1, x_{t-1}, \ldots, x_{t-p})'$ and $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_p)'$. The observed series can be expressed as a multiple linear regression model given by

$$y_t = \delta_t\beta_t + \boldsymbol{\phi}'\boldsymbol{X}_t + a_t \qquad t = p+1, \ldots, n. \tag{2.3}$$

We assume that the outlier indicator δ_t and the outlier magnitude β_t are independent and distributed as $Bernoulli(\alpha)$ and $\mathcal{N}(0, \tau^2)$ respectively for all t, where α and τ^2 are hyperparameters. Therefore, the prior probability of being contaminated by an outlier is the same for all observations, namely $P(\delta_t = 1) = \alpha$, for $t = p+1, \ldots, n$. The prior distribution for the contamination parameter α is $Beta(\gamma_1, \gamma_2)$, with expectation $E(\alpha) = \gamma_1/(\gamma_1 + \gamma_2)$.

Abraham and Box (1979) obtained the posterior distributions for the parameters in model (2.3), under Jeffrey's reference prior distribution for $(\boldsymbol{\phi}, \sigma^2)$. In this case the conditional posterior distribution of $\boldsymbol{\phi}$, given any outlier configuration $\boldsymbol{\delta}_r$, is a multivariate-t distribution, where $\boldsymbol{\delta}_r$ is one of the 2^{n-p} possible outlier configurations. Then the posterior distribution of $\boldsymbol{\phi}$ is a mixture of 2^{n-p} multivariate-t distributions:

$$P(\boldsymbol{\phi} \mid \boldsymbol{y}) = \sum w_r P(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{\delta}_r), \tag{2.4}$$

where the summation is over all 2^{n-p} possible outlier configurations and the weight $w_r = P(\boldsymbol{\delta}_r \mid \boldsymbol{y})$. For such a model, we can identify the outliers using the posterior marginals of elements of $\boldsymbol{\delta}$,

$$p_t = P(\delta_t = 1 \mid \boldsymbol{y}) = \sum P(\boldsymbol{\delta}_{r_t} \mid \boldsymbol{y}), \qquad t = p+1, \ldots, n, \tag{2.5}$$

where the summation is now over the 2^{n-p-1} outlier configurations $\boldsymbol{\delta}_{r_t}$ with $\delta_t = 1$ (the posterior probabilities $P(\boldsymbol{\delta}_{r_t} \mid \boldsymbol{y})$ are easy to compute). The posterior distributions of the outlier magnitudes are mixtures of Student-t distributions:

$$P(\beta_t \mid \boldsymbol{y}) = \sum w_r P(\beta_t \mid \boldsymbol{y}, \boldsymbol{\delta}_r), \qquad t = p+1, \ldots, n. \tag{2.6}$$

Therefore, it is possible to derive the posterior probabilities for the parameters in model (2.3). In practice, however, the computation is very intensive, even when the sample size is small. Since these probabilities are mixtures of 2^{n-p} or 2^{n-p-1} distributions. The approach becomes infeasible when the sample size is moderate or large and some alternative approach is needed. One alternative is to use MCMC methods.

2.2. The standard Gibbs sampling and its difficulties

McCulloch and Tsay (1994) proposed to compute the posterior distributions (2.4) − (2.6) by Gibbs sampling. The procedure requires full conditional posterior distributions of each parameter in model (2.3) given all the other parameters. Barnett, Kohn and Sheather (1996) generalized the model to include innovative outliers and order selection. They used MCMC methods with Metropolis-Hasting and Gibbs Sampling algorithms. For ease of reference, we summarize conditional posterior distributions, obtained first by McCulloch and Tsay (1994), with conjugate prior distributions for $\boldsymbol{\phi}$ and σ_a^2. We use non-informative priors for these parameters. Note that, as the priors for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are proper, even if we assume improper priors for $(\boldsymbol{\phi}, \sigma_a^2)$ the joint posterior is proper.

If the hyperparameters γ_1, γ_2 and τ^2 are known, the conditional posterior distribution of the AR parameter vector $\boldsymbol{\phi}$ is multivariate normal $\mathcal{N}_{p+1}(\boldsymbol{\phi}^\star, \sigma_a^2 \boldsymbol{\Omega}_\phi)$,

where the mean vector and the covariance matrix are

$$\boldsymbol{\phi}^{\star} = \boldsymbol{\Omega}_{\phi} \sum_{t=p+1}^{n} \boldsymbol{X}_t x_t, \quad \text{and} \quad \boldsymbol{\Omega}_{\phi} = \Big(\sum_{t=p+1}^{n} \boldsymbol{X}_t \boldsymbol{X}_t' \Big)^{-1}.$$

The conditional posterior distribution of the innovational variance σ_a^2 is *Inverted-Gamma*$((n-p)/2, (\sum_{t=p+1}^{n} a_t^2)/2)$.

The conditional posterior distribution of α depends only on the vector $\boldsymbol{\delta}$, it is $Beta(\gamma_1 + \sum \delta_t, \gamma_2 + (n-p) - \sum \delta_t))$. The conditional posterior mean of α can then be expressed as a linear combination of the prior mean and the sample mean $\bar{\delta}$ of the data: $E(\alpha \mid \boldsymbol{\delta}) = \omega E(\alpha) + (1-\omega)\bar{\delta}$, where $\omega = (\gamma_1+\gamma_2)/(\gamma_1+\gamma_2+n-p)$.

The conditional posterior distribution of δ_j, $j = p+1, \ldots, n$, is Bernoulli with probability

$$P(\delta_j = 1 \mid \boldsymbol{y}, \boldsymbol{\phi}, \sigma_a^2, \boldsymbol{\delta}_{(j)}, \boldsymbol{\beta}, \alpha) = \Big[1 + \frac{(1-\alpha)}{\alpha} B_{10}(j)\Big]^{-1}, \tag{2.7}$$

where $\boldsymbol{\delta}_{(j)}$ is obtained from $\boldsymbol{\delta}$ by eliminating the element δ_j and B_{10} is the Bayes factor. The logarithm of the Bayes factor B_{10} is

$$\log B_{10}(j) = \frac{1}{2\sigma_a^2}\Big[\sum_{t=j}^{T_j} e_t(1)^2 - \sum_{t=j}^{T_j} e_t(0)^2\Big], \tag{2.8}$$

where $T_j = \min(n, j+p)$, $e_t(\delta_j) = \tilde{x}_t - \phi_0 - \sum_{i=1}^{p} \phi_i \tilde{x}_{t-i}$, and $\tilde{x}_t = y_t - \delta_t \beta_t$ is the residual at time t when the series is corrected by the identified outliers in $\boldsymbol{\delta}_{(j)}$ and δ_j. It is easy to see that $e_t(1) = e_t(0) + \pi_{t-j}\beta_j$, where $\pi_0 = -1$ and $\pi_j = \phi_j$ for $j = 1, \ldots, p$.

The probability (2.7) has a simple interpretation. The two hypotheses $\delta_j = 1$ (y_j is contaminated by an outlier) and $\delta_j = 0$ (y_j is outlier free), given the data, only affect the residuals $e_j, \ldots, e_{T_j}$. Assuming parameters are known, we can judge the likelihoods of these hypotheses by (a) computing the residuals $e_t(1)$ for $t = j, \ldots, T_j$; (b) computing the residuals $e_t(0)$; and (c) comparing the two sets of residuals. The Bayes factor is the usual way of comparing the likelihoods of the two hypotheses. Since the residuals are one-step-ahead prediction errors, (2.8) compares the sum of prediction errors in the periods $j, j+1, \ldots, T_j$ when the forecasts are evaluated under the hypothesis $\delta_j = 1$ and under the hypothesis $\delta_j = 0$. This is equivalent to the Chow test (1960) for structural changes when the variance is known.

The conditional posterior distributions of the outlier magnitudes β_j, for $j = p+1, \ldots, n$ are $\mathcal{N}(\delta_j \beta_j^{\star}, \sigma_j^2)$, where

$$\beta_j^{\star} = \frac{\sigma_j^2}{\sigma_a^2}\Big[e_j(0) - \phi_1 e_{j+1}(0) - \cdots - \phi_{T_j - j} e_{T_j}(0)\Big] \tag{2.9}$$

and $\sigma_j^2 = \tau^2\sigma_a^2/(\tau^2\nu_{T_j-j}^2\delta_j + \sigma_a^2)$, with $\nu_{T_j-j}^2 = (1 + \phi_1^2 + \cdots + \phi_{T_j-j}^2)$.

When y_j is identified as an outlier and there is no prior information about the outlier magnitude ($\tau^2 \to \infty$), the conditional posterior mean of β_j tends to $\hat{\beta}_j = \nu_{T_j-j}^{-2}[e_j(0) - \phi_1 e_{j+1}(0) - \cdots \phi_{T_j-j}e_{T_j}(0)]$, which is the least squares estimate when the parameters are known (Chang, Tiao and Chen (1988)); the conditional posterior variance of β_j tends to the variance of the estimate $\hat{\beta}_j$. The conditional posterior mean in (2.9) can also be seen as a linear combination of the prior mean and the outlier magnitude estimated from the data. The magnitude estimate is the difference between the observation y_j and the conditional expectation of y_j given all the data, $\hat{y}_{j|n}$, which is the linear predictor of y_j that minimizes the mean squared error. Then (2.9) can be expressed as

$$\beta_j^\star = \frac{\tau^2\nu_{T_j-j}^2}{\tau^2\nu_{T_j-j}^2 + \sigma_a^2}(y_j - \hat{y}_{j|n}) + \frac{\sigma_a^2}{\tau^2\nu_{T_j-j}^2 + \sigma_a^2}\,\beta_0, \tag{2.10}$$

where β_0 is the prior mean of β_j (zero in this paper). For the AR(p) model under study, the optimal linear predictor $\hat{y}_{j|n}$ is a combination of the past and future p values of y_j,

$$\hat{y}_{j|n} = \phi_0\nu_{T_j-j}^{-2}\tilde{\pi}_{T_j-j} - \nu_{T_j-j}^{-2}\Big(\sum_{i=1}^{p}\sum_{t=0}^{T_j-j-i}\pi_t\pi_{t+i}x_{j-i} + \sum_{i=1}^{T_j-j}\sum_{t=0}^{T_j-j-i}\pi_t\pi_{t+i}x_{j+i}\Big), \tag{2.11}$$

where $\tilde{\pi}_t = 1 - \phi_1 - \cdots - \phi_t$ for $t \le p$. Using the truncated autoregressive polynomial $\pi_{T_j-j}(B) = (1 - \pi_1 B - \cdots - \pi_{T_j-j}B^{T_j-j})$ and the "truncated" variance, $\nu_{T_j-j}^2 = (1 + \pi_1^2 + \cdots + \pi_{T_j-j}^2)$ of the dual process

$$x_t^{\mathrm{D}} = \phi_0\tilde{\pi}_p + a_t - \phi_1 a_{t-1} - \cdots - \phi_p a_{t-p}, \tag{2.12}$$

the filter (2.11) can be written as a function of the "truncated" autocorrelation generating function $\rho_{T_j-j}^{\mathrm{D}}(B) = \nu_{T_j-j}^{-2}\pi_p(B)\pi_{T_j-j}(B^{-1})$ of the dual process $\hat{y}_{j|n} = \phi_0\nu_{T_j-j}^{-2}\tilde{\pi}_{T_j-j} - [1 - \rho_{T_j-j}^{\mathrm{D}}(B)]x_j$.

We may use the above results to draw a sample of a Markov chain using Gibbs sampling. This Markov chain converges to the joint posterior distribution of the parameters. When the number of iterations is sufficiently large, the Gibbs draw can be regarded as a sample from the joint posterior distribution. These draws are easy to obtain because they are from well-known probability distributions. However, as shown by the simple example in 1.2, such a Gibbs sampling procedure may fare poorly when additive outliers appear in patches.

To understand the situation more fully, consider the simplest situation in which the time series follows an AR(1) model with mean zero and there exists a patch of three additive outliers at time $T-1$, T, $T+1$. To simplify the

analysis, we first assume that the three outliers have the same size, $\beta_t = \beta$ for $t = T-1, T, T+1$, and the AR parameter is known. Checking if an observation is contaminated by an additive outlier amounts to comparing the observed value with its optimal interpolator derived using the model and the rest of the data. For an AR(1) process the optimal interpolator (2.11) is $\hat{y}_{t|n} = \phi(1+\phi^2)^{-1}(y_{t-1}+y_{t+1})$. The first outlier in the patch is tested by comparing $y_{T-1} = x_{T-1} + \beta_{T-1}$ with $\hat{y}_{T-1|n} = \hat{x}_{T-1|n} + \phi(1+\phi^2)^{-1}\beta_T$. It is detected if β_T is sufficiently large, but its size will be underestimated. For instance, if ϕ is close to unity the estimated size will only be half the true size. For the second outlier, we compare $y_T = x_T + \beta_T$ with $\hat{y}_{T|n} = \hat{x}_{T|n} + \phi(1+\phi^2)^{-1}(\beta_{T-1} + \beta_{T+1}) = \hat{x}_{T|n} + (1+\phi^2)^{-1}(2\phi\beta_T)$. If ϕ is close to one the outlier cannot be easily detected, because the difference between the optimal interpolator and the observed value is $x_T - \hat{x}_{T|n} + (1-\phi)^2(1+\phi^2)^{-1}\beta$, small when ϕ is close to unity. Note that in practice the masking effect is likely to remain even if we correctly identify an outlier at $T-1$ and adjust y_{T-1} accordingly, because, as mentioned above, the outlier size at $T-1$ is underestimated. In short, if ϕ is close to unity the middle outlier at time T is hard to detect. The detection of the outlier at time $T+1$ is also difficult and its size is underestimated.

Next consider the case that the AR parameter is estimated from the sample. Without outliers the least squares estimate of the AR parameter is $\hat{\phi}_0 = \sum x_t x_{t-1}(\sum x_t^2)^{-1} = r_x(1)$, the lag-1 sample autocorrelation. Suppose the series is contaminated by a patch of $2k+1$ additive outliers at times $T-k, \ldots, T, \ldots, T+k$, of sizes $\beta_t = \beta_t^o s_x$, where $s_x^2 = \sum x_t^2/n$ is the sample variance of the outlier-free series. In this case, the least squares estimate of the AR coefficient based on the observed time series y_t is given, dropping terms of order $o(n^{-1})$, by

$$\hat{\phi_y} = \frac{r_x(1) + n^{-1}\sum_{-k}^{k}\beta_{T+j}^o(\tilde{x}_{T+j-1} + \tilde{x}_{T+j+1}) + n^{-1}\sum_{-k}^{k}\beta_{T+j}^o\beta_{T+j-1}^o}{1 + 2n^{-1}\sum_{-k}^{k}\beta_{T+j}^o\tilde{x}_{T+j} + n^{-1}\sum_{-k}^{k}(\beta_{T+j}^o)^2},$$

where $\tilde{x}_t = x_t/s_x$. If the outliers are large, with the same sign and similar sizes, then for a fixed sample size n, $\hat{\phi_y}$ will approach unity as the outlier sizes increase. This makes the identification of outliers difficult.

Note that characteristics of the outlier patch will appear clearly if the length of the patch is known and one interpolates the whole patch using observations *before* and *after* the patch. This is the main idea of the adaptive Gibbs Sampling proposed in the next section. For the standard outlier detection procedure using Gibbs sampling, proper "block" interpolation can only occur if one uses ideal initial values that identify each point of the outlier patch as an outlier.

Figure 2 shows the evolution of the masking problem when the Gibbs sampling iteration starts for the simulated data in (1.2). At each iteration, the Gibbs sampler checks if an observation is an outlier by comparing the observed

value with its optimal interpolator using (2.8) and (2.10). The outlier sizes for β_{37} to β_{42} generated in each iteration are represented in Figure 2 as continuous sequences —grey shadows represent iterations where the data are identified as outliers. The outliers at $t = 38$ and 41 are identified and corrected in few iterations, but their sizes are underestimated. The central outliers, $t = 39$ and 40, are identified in very few iterations and their sizes on these occasions are small. We can see that four outliers are never identified in the same iteration. If we use ideal initial values that assign each point of the outlier patch as an outlier, the sizes obtained at an iteration are the dots in Figure 2 (horizontal lines represent true outlier sizes). At the right side of the graphs, the estimation of the posterior distributions are represented for standard Gibbs sampling (the fill curve) and for the Gibbs sampler with "block" interpolation.

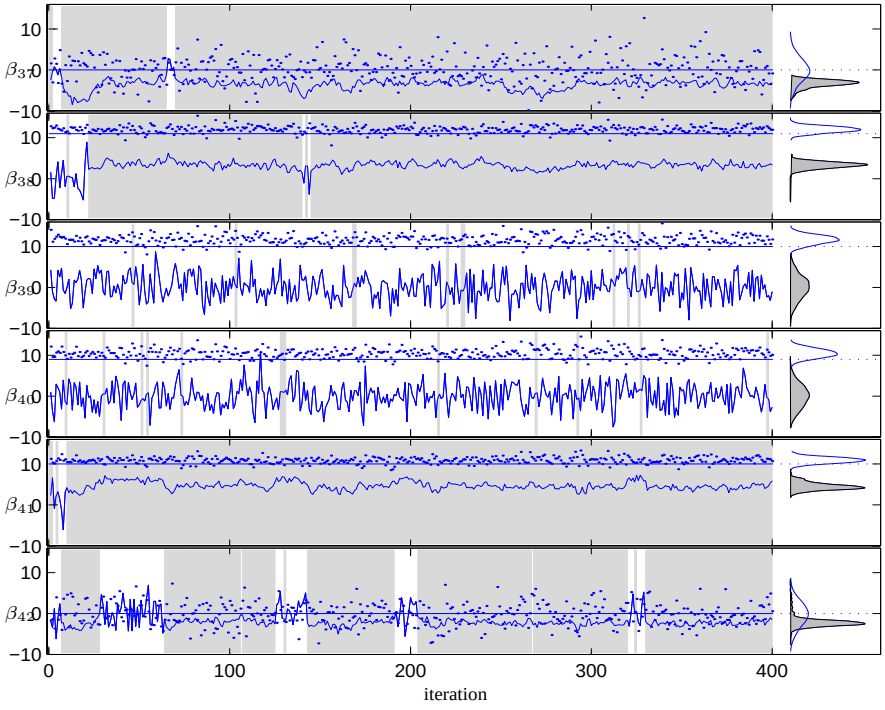


Figure 2. Artificial time series: outlier sizes from β_{37} to β_{42} generated in each iteration. *Continuous sequences* for the standard Gibbs sampler, and *dots* for the Gibbs sampler with "block" interpolation. The grey shadows are used to show the iterations where the data are identified as outliers. The horizontal lines represent the true values.

One can regard these difficulties as a practical convergence problem for the Gibbs sampler. This problem also appears in the regression case with multiple

outliers, as shown by Justel and Peña (1996). High parameter correlations and the large dimension of the parameter space slow down the convergence of the Gibbs sampler (see Hills and Smith (1992)). Note that the dimension of the parameter space is $2n + p + 3$ and, for outliers in a patch, correlations are large among the outlier positions and among the outlier magnitudes. When Gibbs draws are from a joint distribution of highly correlated parameters, movements from one iteration to the next are in the principal components direction of the parameter space instead of parallel to the coordinate axes.

3. Detecting Outlier Patches

Our new procedure consists of two Gibbs runs. In the first run, the standard Gibbs sampling of Section 2 is applied to the data. The results of this Gibbs run are then used to implement a second Gibbs sampling that is adaptive in treating identified outliers and in using block interpolation to reduce possible masking and swamping effects.

In what follows, we divide the detail of the proposed procedure into subsections. The discussion focuses on the second Gibbs run and assumes that results of the first Gibbs run are available. For ease in reference, let $\hat{\boldsymbol{\phi}}^{(s)}$, $\hat{\sigma}_a^{(s)}$, $\hat{\boldsymbol{p}}^{(s)}$, and $\hat{\boldsymbol{\beta}}^{(s)}$ be the posterior means based on the last r iterations of the first Gibbs run which uses s iterations, where the jth element of $\hat{\boldsymbol{p}}^{(s)}$ is $\hat{p}_{p+j}^{(s)}$, the posterior probability that y_{p+j} is contaminated by an outlier.

3.1. Location and joint estimation of outlier patches

The biases in $\hat{\boldsymbol{\beta}}^{(s)}$ induced by the masking effects of multiple outliers come from several sources. Two main sources are (a) drawing values of β_j one by one and (b) the misspecification of the prior mean of β_j, fixed to zero. One-by-one drawing overlooks the dependence between parameters. For an AR(p) process, an additive outlier affects $p + 1$ residuals and the usual interpolation (or filtering) involves p observations before and after the time index of interest. Consequently, an additive outlier affects the conditional posterior distributions of $2p+1$ observations; see (2.8) and (2.9). Chen and Liu (1993) pointed out that estimates of outlier magnitudes computed separately can differ markedly from those obtained from a joint estimation. The situation becomes more serious in the presence of k consecutive additive outliers for which the outliers affect $2p+k$ observations. We make use of the results of the first Gibbs sampler to identify possible locations and block sizes of outlier patches.

The tentative specification of locations and block sizes of outlier patches is done by a forward-backward search using a window around the outliers identified by the first Gibbs run. Let c_1 be a critical value between 0 and 1 used to identify

potential outliers. An observation whose posterior probability of being an outlier exceeds c_1 is classified as an "identified" outlier. More specifically, y_j is identified as an outlier if $\hat{p}_j^{(s)} > c_1$. Typically we use $c_1 = 0.5$. Let $\{t_1, \ldots, t_m\}$ be the collection of time indexes of outliers identified by the first Gibbs run.

Consider patch size. We select another critical value c_2, $c_2 \leq c_1$, to specify the beginning and end points of a "potential" outlier patch associated with an identified outlier. In addition, because the length of an outlier patch cannot be too large relatively to the sample size, we select a window of length $2hp$ to search for the boundary points of a possible outlier patch. For example, consider an "identified" outlier y_{t_i}. First, we check the hp observations *before* y_{t_i} and compare their posterior probabilities $\hat{p}_j^{(s)}$ with c_2. Any point within the window with $\hat{p}_j^{(s)} > c_2$ is regarded as a possible beginning point of an outlier patch associated with y_{t_i}. We then select the farthest point from y_{t_i} as the beginning point of the outlier patch. Denote the point by $y_{t_i-k_i}$. Second, do the same for the hp observations *after* y_{t_i} and select the farthest point from y_{t_i} with $\hat{p}_j^{(s)} > c_2$ as the end point of the outlier patch. Denote the end point by $y_{t_i+v_i}$. Combine the two blocks to form a tentative candidate for an outlier patch associated with y_{t_i}, denoted by $(y_{t_i-k_i}, \ldots, y_{t_i+v_i})$.

Finally, consider jointly all the identified outlier patches for further refinement. Overlapping or consecutive patches should be merged to form a larger patch; if the total number of outliers is greater than $n/2$, where n is the sample size, increase c_2 and re-specify possible outlier patches; if increasing c_2 cannot sufficiently reduce the total number of outliers, choose a smaller h and re-specify outlier patches.

With outlier patches tentatively specified, draw Gibbs samples jointly within a patch. Suppose that a patch of k outliers starting at time index j is identified. Let $\boldsymbol{\delta}_{j,k} = (\delta_j, \ldots, \delta_{j+k-1})'$ and $\boldsymbol{\beta}_{j,k} = (\beta_j, \ldots, \beta_{j+k-1})'$ be the vectors of outlier indicators and magnitudes, respectively, for the patch. To complete the sampling scheme we need the conditional posterior distributions of $\boldsymbol{\delta}_{j,k}$ and $\boldsymbol{\beta}_{j,k}$, given the others. We give these distributions in the next theorem, derivations are in the Appendix.

Theorem 1. *Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be a vector of observations according to (2.3), with no outliers in the first p data points. Assume $\delta_t \sim$ Bernoulli(α), $t = p+1, \ldots, n$, and*

$$P(\boldsymbol{\phi}, \sigma_a^2, \alpha, \boldsymbol{\beta}) \propto \sigma_a^{-2} \alpha^{\gamma_1 - 1} (1-\alpha)^{\gamma_2 - 1} \exp\left(-\frac{1}{2\tau^2} \sum_{t=p+1}^{n} \beta_t^2 \right),$$

where the parameters γ_1, γ_2 and τ^2 are known. Let $e_t(\boldsymbol{\delta}_{j,k}) = x_t - \phi_0 - \sum_{i=1}^{p} \phi_i x_{t-i}$ be the residual at time t when the series is adjusted for all identified outliers

not in the interval $[j, j+k-1]$ *and the outliers identified in* $\boldsymbol{\delta}_{j,k}$*, with* $T_{j,k} = \min\{n, j+k+p-1\}$*. Then the following hold.*

a) *The conditional posterior probability of a block configuration* $\boldsymbol{\delta}_{j,k}$*, given the sample and the other parameters, is*

$$p_{\delta_{j,k}} = C\ \alpha^{\boldsymbol{s}_{j,k}}(1-\alpha)^{k-\boldsymbol{s}_{j,k}}\ \exp\Big(-\frac{1}{2\sigma_a^2}\sum_{t=j}^{T_{j,k}} e_t(\boldsymbol{\delta}_{j,k})^2\Big), \tag{3.13}$$

where $\boldsymbol{s}_{j,k} = \sum_{t=j}^{j+k-1}\delta_t$*, and* C *is a normalization constant so that the total probability of the* 2^k *possible configurations of* $\boldsymbol{\delta}_{j,k}$ *is one.*

b) *The conditional posterior distribution of* $\boldsymbol{\beta}_{j,k}$ *given the sample and other parameters is* $\mathcal{N}_k\left(\boldsymbol{\beta}^{\star}_{j,k}, \boldsymbol{\Omega}_{j,k}\right)$*,*

$$\boldsymbol{\beta}^{\star}_{j,k} = \boldsymbol{\Omega}_{j,k}\Big(-\frac{1}{\sigma_a^2}\sum_{t=j}^{T_{j,k}} e_t(\mathbf{0})\boldsymbol{D}_{j,k}\boldsymbol{\Pi}_{t-j} + \frac{1}{\tau^2}\boldsymbol{\beta}_0\Big), \tag{3.14}$$

$$\boldsymbol{\Omega}_{j,k} = \Big(\boldsymbol{D}_{j,k}\Big(\frac{1}{\sigma_a^2}\sum_{t=j}^{T_{j,k}}\boldsymbol{\Pi}_{t-j}\boldsymbol{\Pi}'_{t-j}\Big)\boldsymbol{D}_{j,k} + \frac{1}{\tau^2}\boldsymbol{I}\Big)^{-1}, \tag{3.15}$$

where $\boldsymbol{D}_{j,k}$ *is a* $k\times k$ *diagonal matrix with elements* $\delta_j, \ldots, \delta_{j+k-1}$*, and* $\boldsymbol{\Pi}_t = (\pi_t, \pi_{t-1}, \ldots, \pi_{t-k+1})'$ *is a* $k\times 1$ *vector, with* $\pi_0 = -1$*,* $\pi_i = \phi_i$ *for* $i = 1, \ldots, p$*, and* $\pi_i = 0$ *for* $i < 0$ *or* $i > p$*.*

After computing the probabilities (3.13) for all 2^k possible configurations for the block $\boldsymbol{\delta}_{j,k}$, the outlying status of each observation in the outlier patch will be classified separately. Another possibility, suggested by a referee, is to engineer some Metropolis-Hasting moves by using Theorem 1. An advantage of our procedure is that we can generate large block configurations from (3.13) without computing C, but an optimal criterion (in the sense of acceptance rate and moving) should be found. This possibility will be explored in future work.

Let $\boldsymbol{W}_1 = \sigma_a^{-2}\boldsymbol{\Omega}_{j,k}(\boldsymbol{D}_{j,k}\sum_{t=j}^{T_{j,k}}\boldsymbol{\Pi}_{t-j}\boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k})$ and $\boldsymbol{W}_2 = \tau^{-2}\boldsymbol{\Omega}_{j,k}$. Then $\boldsymbol{\beta}^{\star}_{j,k}$ can be written as $\boldsymbol{\beta}^{\star}_{j,k} = \boldsymbol{W}_1\tilde{\boldsymbol{\beta}}_{j,k} + \boldsymbol{W}_2\boldsymbol{\beta}_0$, where $\boldsymbol{W}_1 + \boldsymbol{W}_2 = \boldsymbol{I}$, implying that the mean of the conditional posterior distribution of $\boldsymbol{\beta}_{j,k}$ is a linear combination of the prior mean vector $\boldsymbol{\beta}_0$ and the least squares estimate (or the maximum likelihood estimate) of the outlier magnitudes for an outlier patch

$$\tilde{\boldsymbol{\beta}}_{j,k} = \Big(\boldsymbol{D}_{j,k}\sum_{t=j}^{T_{j,k}}\boldsymbol{\Pi}_{t-j}\boldsymbol{\Pi}'_{t-j}\boldsymbol{D}_{j,k}\Big)^{-1}\Big(-\sum_{t=j}^{T_{j,k}} e_t(\mathbf{0})\boldsymbol{D}_{j,k}\boldsymbol{\Pi}_{j,k}\Big). \tag{3.16}$$

Peña and Maravall (1991) proved that, when $\delta_t = 1$, the estimate in (3.16) is equivalent to the vector of differences between the observations $(y_j, \ldots, y_{j+k-1})$

and the predictions $\hat{y}_t = E(y_t \mid y_1, \ldots, y_{j-1}, y_{j+k}, \ldots, y_n)$ for $t = j, \ldots, j+k-1$. The matrix $\mathbf{\Pi} = \sum_{t=j}^{T_{j,k}} \mathbf{\Pi}_{t-j}\mathbf{\Pi}'_{t-j}$ is the $k \times k$ submatrix of the "truncated" autocovariance generating matrix of the dual process in (2.12). Specifically,

$$\mathbf{\Pi} = \begin{pmatrix} \nu^2_{T_{j,k}-j} & \gamma^{\text{D}}_{1,T_{j,k}-j-1} & \cdots & \gamma^{\text{D}}_{k-1,T_{j,k}-j-k+1} \\ \gamma^{\text{D}}_{-1,T_{j,k}-j} & \nu^2_{T_{j,k}-j-1} & \cdots & \gamma^{\text{D}}_{k-2,T_{j,k}-j-k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^{\text{D}}_{-k+1,T_{j,k}-j} & \gamma^{\text{D}}_{-k+2,T_{j,k}-1} & \cdots & \nu^2_{T_{j,k}-j-k+1} \end{pmatrix},$$

where $\gamma^{\text{D}}_{i,j} = \nu^2_j \rho^{\text{D}}_{i,j}$, ν^2_j is the "truncated" variance of the dual process and $\rho^{\text{D}}_{i,j}$ is the coefficient of B^i in the "truncated" autocorrelation generating function of the dual process, i.e., $\rho_j{}^{\text{D}}(B) = \nu_j^{-2}\pi_p(B)\pi_j(B^{-1})$.

3.2. The second Gibbs sampling

We discuss the second adaptive Gibbs run of the proposed procedure. The results of the first Gibbs run provide useful information to start the second Gibbs sampling and to specify prior distributions of the parameters. The starting values of δ_t are as follows: $\delta_t^{(0)} = 1$ if $\hat{p}_t^{(s)} > 0.5$, i.e., if y_t belongs to an identified outlier patch; otherwise, $\delta_t^{(0)} = 0$. The prior distributions of β_t are as follows.

a) If y_t is identified as an isolated outlier the prior distribution of β_t is $\mathcal{N}(\hat{\beta}_t^{(s)}, \tau^2)$, where $\hat{\beta}_t^{(s)}$ is the Gibbs estimate of β_t from the first Gibbs run.
b) If y_t belongs to an outlier patch the prior distribution of β_t is $\mathcal{N}(\tilde{\beta}_t^{(s)}, \tau^2)$, where $\tilde{\beta}_t^{(s)}$ is the conditional posterior mean given in (3.16).
c) If y_t does not belong to any outlier patch, and is not an isolated outlier, then the prior distribution of β_t is $\mathcal{N}(0, \tau^2)$.

For each outlier patch, the results of Theorem 1 are used to draw $\boldsymbol{\delta}_{j,k}$ and $\boldsymbol{\beta}_{j,k}$ in the second Gibbs sampling. The second Gibbs sampling is also run for s iterations, but only the results of the last r iterations are used to make inference. The number s can be determined by any sequential method proposed in the literature to monitor the convergence of Gibbs sampling. We use a method that can be easily implemented, based on comparing the estimates of outlying probability for each data point, computed with non-overlapping segments of samples. Specifically, after a burn-in period of $b = 5,000$ iterations, we assume convergence has been achieved if the standard test for the equality of two proportions is not rejected. Thus, calling $\hat{p}_t^{(s)}$ the probability that y_t is an outlier computed with the iterations from $s-r+1$ to s, we assume convergence if for all $t = p+1, \ldots, n$ the differences $\left|\hat{p}_t^{(s)} - \hat{p}_t^{(s-r)}\right|$ are smaller than $\epsilon = 3\sqrt{0.5^2/s_r}$.

An alternative procedure for handling outlier patches is to use the ideas of Bruce and Martin (1989). This procedure involves two steps. First, select a

positive integer k in the interval $[1, n/2]$ as the maximum length of a outlier patch. Second, start the Gibbs sampler with $n-k-p$ parallel trials. In the jth trial, for $j = 1, \ldots, n-k-p$, the points at $t = p+j$ to $p+k+j$ are assigned initially as outliers. For other data points, use the usual method to assign initial values. In application, one can use several different k values. However, such a procedure requires intensive computation, especially when n is large.

4. Applications

Here we re-analyze the simulated time series of Section 1 and then consider some real data. We compare the results of the usual Gibbs sampling, referred to as standard Gibbs sampling, with those of the adaptive Gibbs sampling to see the efficacy of the latter algorithm. The example demonstrates the applicability and effectiveness of the adaptive Gibbs sampling, and it shows that patches of outliers occur in applications.

Table 1. Outlier magnitudes: true values and estimates obtained by the standard and adaptive Gibbs samplings.

Parameter	β_{27}	β_{37}	β_{38}	β_{39}	β_{40}	β_{41}	β_{42}
True Value	-3	0	11	10	9	10	0
Standard GS	-3.03	-3.42	3.31	0.05	-0.05	4.51	-2.09
Adaptive GS	-3.06	-0.09	11.97	11.63	10.43	10.91	-0.23

4.1. Simulated data revisited

As shown in Figure 1, standard Gibbs sampling can easily detect the isolated outlier at $t = 27$ of the simulated AR(3) example, but it has difficulty with the outlier patch in the period $38 - 41$. For the adaptive Gibbs sampling, we choose hyperparameters $\gamma_1 = 5$, $\gamma_2 = 95$ and $\tau = 3$, implying that the contamination parameter has a prior mean $\alpha_0 = 0.05$, and the prior standard deviation of β_t is three times the residual standard deviation. Using $\epsilon = 0.047$ to monitor convergence, we obtained $s = 26,000$ iterations for the first Gibbs sampling and $s = 7,000$ iterations for the second, adaptive Gibbs sampling. All parameter estimates reported are the sample means of the last $r = 1,000$ iterations. For specifying the location of an outlier patch, we chose $c_1 = 0.5$, $c_2 = 0.3$, and the window length $2p$ to search for boundary points of possible outlier patches, where $p = 3$ is the autoregressive order of the series. Additional checking confirms that results are stable over minor modifications of these parameter values.

The results of the first run are shown in Figure 1 and summarized in Table 1. As before, the procedure indicates a possible patch of outliers from $t = 37$ to 42. In the second run the initial conditions and the prior distributions are specified by the proposed adaptive procedure. The posterior probability of outlier for each

data point, $\hat{p}_t^{(s)}$, is shown in Figure 3-a, and the posterior mean of the outlier sizes is shown in Figure 3-b. Adaptive Gibbs sampling successfully specifies all outliers, and there are no swamping or masking effects. In Figure 3-c we compare the posterior distributions of the adaptive (shadow area) and the standard (dotted curve) Gibbs sampling for the error variance and the autoregressive parameters; in Table 1 we compare some of the outlier sizes with the true values. One clearly sees the efficacy and added value of the adaptive Gibbs sampling in this way.

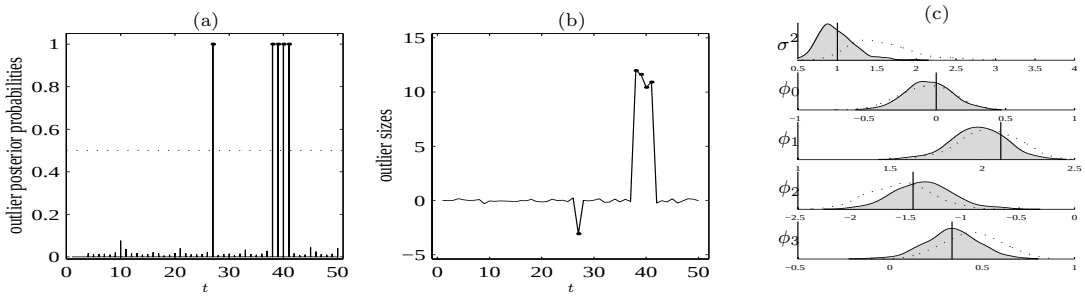


Figure 3. Adaptive Gibbs sampling results with 7,000 iterations for the artificial time series with five outliers: (a) posterior probabilities for each data point to be outlier, (b) posterior mean estimates of the outlier sizes for each data, and (c) kernel estimates of the posterior marginal parameter distributions; the dotted lines are estimates from the first run and vertical lines mark true values.

Figure 4 shows the scatterplots of Gibbs draws of outlier sizes for $t = 37, \ldots, 42$. The right panel is for adaptive Gibbs sampling whereas the left panel is for the standard Gibbs sampling. The plots on the diagonals are histograms of outlier sizes. Adaptive Gibbs sampling exhibits high correlations between sizes of consecutive outliers, in agreement with outlier sizes used. On the other hand, scatterplots of the standard Gibbs sampling do not adequately show correlations between outlier sizes.

Finally, we also ran a more extensive simulation study where the locations for the isolated outlier and the patch were randomly selected. Using the same x_t sequence, we ran standard and adaptive Gibbs sampling for 200 cases with the following results.

1. The isolated outlier was identified by standard and adaptive Gibbs sampling procedures in all cases.
2. The standard Gibbs sampler failed to detect all the outliers in each of the 200 simulations. A typical outcome, as pointed out in Section 3, had the extreme points of the outlier patch and their neighboring points identified as outliers; observations in the middle of the outlier patch were subjected

to masking effects. Occasionally, the standard Gibbs sampler did correctly identify three of the four outliers in the patch. Figure 5-*left* shows a bar plot for the relative frequencies of outlier detection for each data point in the 200 runs. We summarize the relative frequency of identification for each outlier in the patch (bars 1st-4th) and their two neighboring "good" observations.

3. In all simulations, the proposed procedure indicates a possible patch of outliers that includes the true outlier patch.
4. The adaptive Gibbs sampler failed in 13 cases, corresponding to randomly selected patches in the same region of the series. Figure 5-*right* shows the positive performance of the adaptive algorithm in detecting all outliers in the patch.

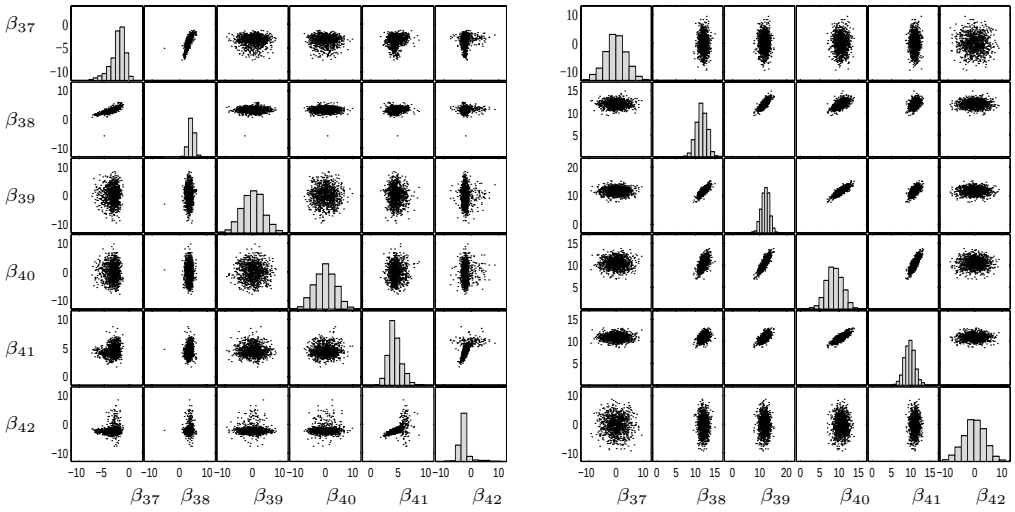


Figure 4. Scatterplots of the standard (*left*) and adaptive (*right*) Gibbs sampler output for β_{37} to β_{42} and the histograms of each magnitude in the diagonal.

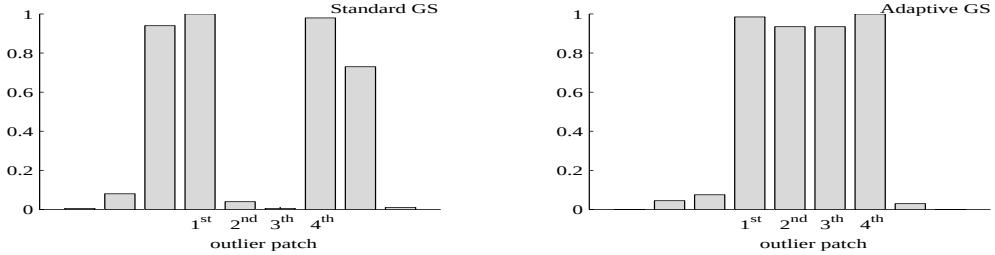


Figure 5. Bar plots for the relative frequencies, in 200 simulations, of outlier detection of each data in the patch (bars 1st-4th), previous and next "good" observations. *Left:* standard Gibbs sampling. *Right:* adaptive Gibbs sampling.

4.2. A real example

Consider the data of monthly U.S. Industry-unfilled orders for radio and TV, in millions of dollars, as studied by Bruce and Martin (1989) among others. We use the logged series from January 1958 to October 1980, and focus on the seasonally adjusted series, where the seasonal component was removed by the well-known X11-ARIMA procedure. The seasonally adjusted series is shown in Figure 6, and can be download from the same web site as the simulated data. An AR(3) model is fit to the data and the estimated parameter values are given in the first row of Table 2. The residual plot of the fit, also shown in Figure 6, indicates possible isolated outliers and outlier patches, especially in the latter part of the series.

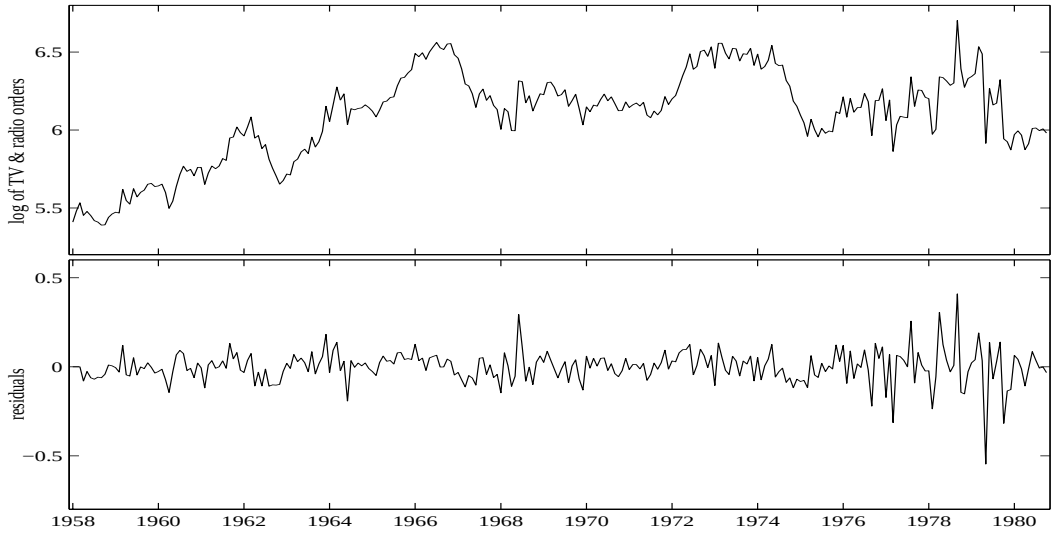


Figure 6. *Top:* Seasonally adjusted series of the logarithm of U.S. Industry-unfilled orders for radio and TV. *Bottom:* Residual plot for the AR(3) model fitted to the series.

Table 2. Estimated parameter values with the initial model and those obtained by the standard and the adaptive Gibbs sampling algorithms.

Parameter	ϕ_0	ϕ_1	ϕ_2	ϕ_3	σ_a
Initial model	0.28	0.61	0.19	0.15	0.091
Standard GS	0.18	0.83	0.19	-0.05	0.062
Adaptive GS	0.18	0.78	0.23	-0.04	0.062

The hyperparameters needed to run the adaptive Gibbs algorithm are set by the same criteria as those of the simulated example: $\gamma_1 = 5$, $\gamma_2 = 95$, $\tau =$

$3\sigma_a = 0.273$. In this particular instance, $r = 1,000$ and the stopping criterion $\epsilon = 0.047$ is achieved by $96,000$ iterations in the first Gibbs run and by $11,000$ iterations in the second. As before, to specify possible outlier patches prior to running the adaptive Gibbs sampling, the window width is set to twice the AR order, $c_1 = 0.5$ and $c_2 = 0.3$. In addition, we assume that the maximum length of an outlier patch is 11 months, just below one year.

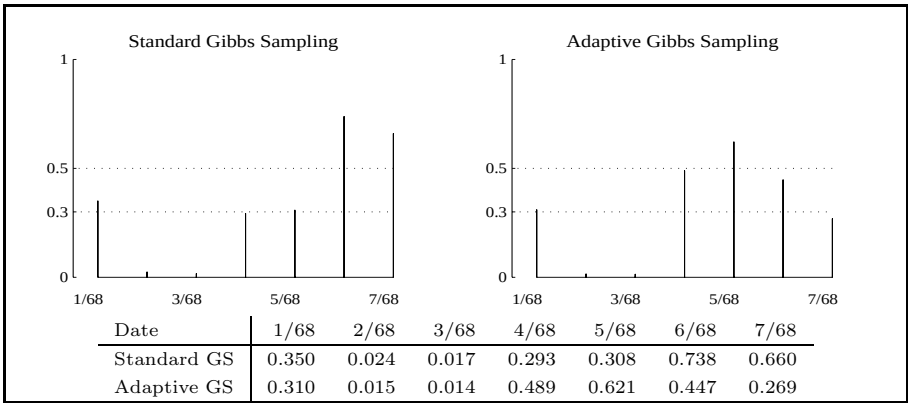
Table 3. Posterior probabilities for each data point to be an outlier by the standard and the adaptive Gibbs sampling algorithms. Estimated outlier sizes by the standard and the adaptive Gibbs sampling algorithms.

	Date	5/68	6/68	7/68	1/73	9/76	1/77	3/77	8/77
Standard GS	Outlier probability	0.31	0.74	0.66	0.54	0.99	0.53	0.99	0.99
	Outlier size	-0.07	0.17	0.12	-0.09	-0.22	-0.08	-0.23	0.21
Adaptive GS	Outlier probability	0.62	0.45	0.27	0.37	0.99	0.24	0.99	1.00
	Outlier size	-0.15	0.14	0.13	-0.06	-0.21	-0.03	-0.20	0.21
	Date	2/78	3/78	9/78	3/79	4/79	5/79	9/79	
Standard GS	Outlier probability	1.00	1.00	1.00	0.51	0.36	1.00	1.00	
	Outlier size	-0.26	-0.27	0.37	0.09	0.06	-0.41	0.26	
Adaptive GS	Outlier probability	1.00	1.00	1.00	0.92	0.89	1.00	1.00	
	Outlier size	-0.28	-0.28	0.37	0.24	0.23	-0.28	0.32	

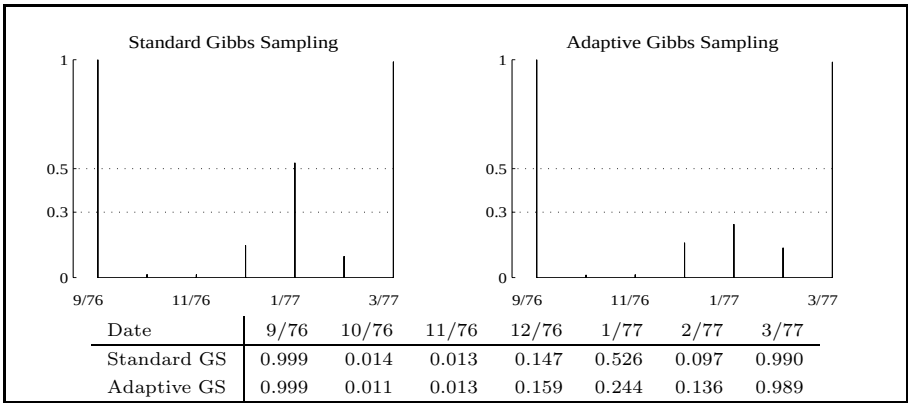
Using 0.5 as the cut-off posterior probability to identify outliers, we summarize the results of standard and adaptive Gibbs sampling in Table 3. The standard Gibbs algorithm identifies 13 data points as outliers. Nine isolated outliers and two outlier patches both of length 2. The two outlier patches are 6-7/1968 and 2-3/1978. The outlier posterior probability for each data point is shown in Figure 7. On the other hand, the second, adaptive Gibbs sampling specifies 11 data points as outliers —six isolated outliers, and two outlier patches of length 2 and 3 at 2-3/1978 and 3-5/1979, respectively. The outlier posterior probabilities based on adaptive Gibbs sampling are also presented in Figure 7. Finally, Table 4 presents outlier posterior probabilities for each data point in the detected outlier patches, for both the standard and adaptive Gibbs samplings. For the possible outlier patch from January to July 1968, the two algorithms show different results: standard Gibbs sampling identifies the patch 6-7/1968 as outliers; adaptive Gibbs sampling detects an isolated outlier at 5/68. For the possible outlier patch from September 1976 to March 1977, the standard algorithm detects an isolated outlier at 1/77, while the adaptive algorithm does not detect any outlier within the patch. For the possible outlier patch from March to September 1979, the standard algorithm identifies two isolated outliers in April and September. On the other hand, the adaptive algorithm substantially

increases the outlying posterior probabilities for March and April of 1979 and, hence, changes an isolated outlier into a patch of three outliers. The isolated outlier in September remains unchanged. Based on the estimated outlier sizes in Table 3, the standard Gibbs algorithm seems to encounter severe masking effects for March and April of 1979.

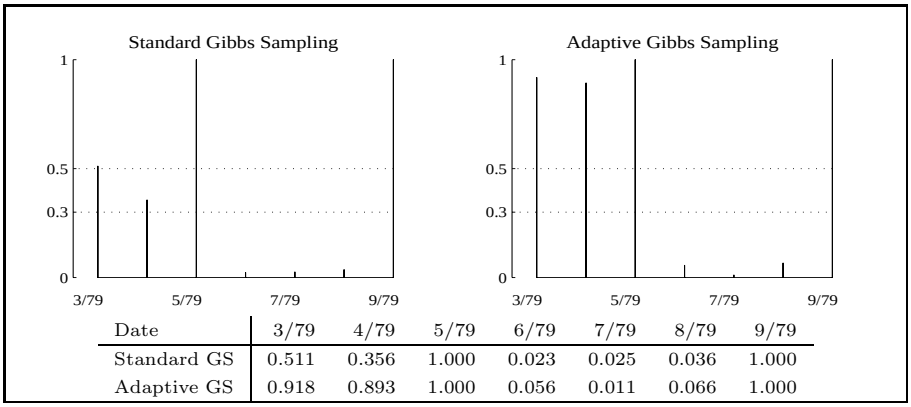
Table 4. Posterior outlier probabilities for each data point in the three larger possible outlier patches.



Date	1/68	2/68	3/68	4/68	5/68	6/68	7/68
Standard GS	0.350	0.024	0.017	0.293	0.308	0.738	0.660
Adaptive GS	0.310	0.015	0.014	0.489	0.621	0.447	0.269



Date	9/76	10/76	11/76	12/76	1/77	2/77	3/77
Standard GS	0.999	0.014	0.013	0.147	0.526	0.097	0.990
Adaptive GS	0.999	0.011	0.013	0.159	0.244	0.136	0.989



Date	3/79	4/79	5/79	6/79	7/79	8/79	9/79
Standard GS	0.511	0.356	1.000	0.023	0.025	0.036	1.000
Adaptive GS	0.918	0.893	1.000	0.056	0.011	0.066	1.000

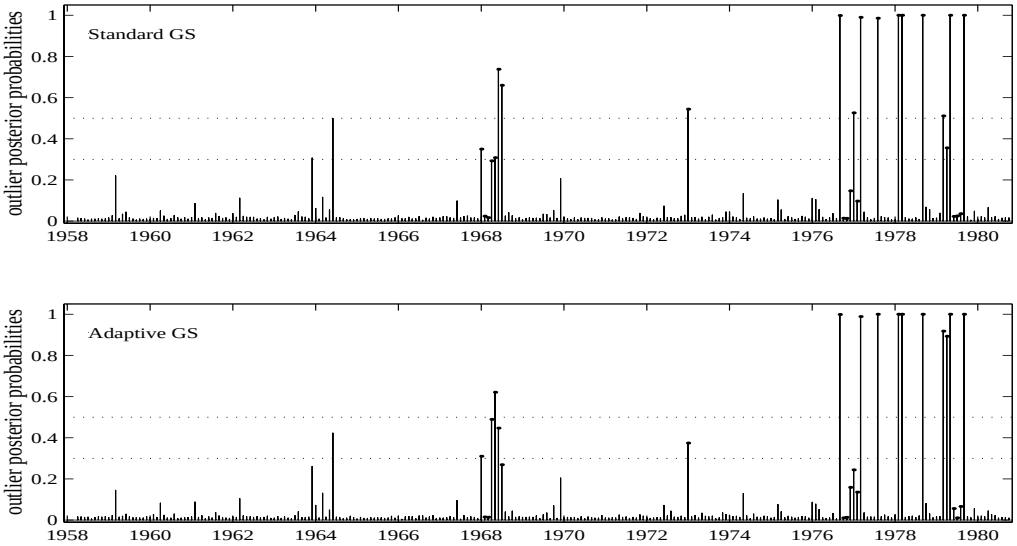


Figure 7. Posterior probability for each data point to be outlier with the standard (*top*) and the adaptive (*bottom*) Gibbs sampling.

Figure 8 shows the time plot of posterior means of residuals obtained by adaptive Gibbs sampling and should be compared with the residuals we obtained without outlier adjustment in Figure 6.

The results of this example demonstrate that (a) outlier patches occur frequently in practice and (b) the standard Gibbs sampling for outlier detection may encounter severe masking and swamping effects, while the adaptive Gibbs sampling performs reasonably well.

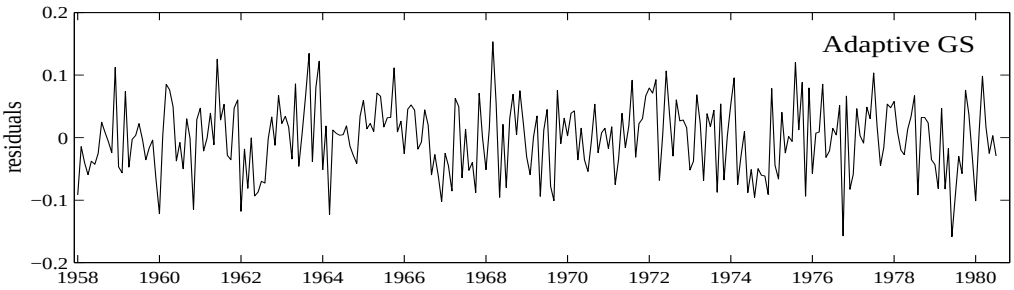


Figure 8. Mean residuals with the adaptive Gibbs sampler.

Acknowledgements

The research of Ana Justel was financed by the European Commission in the Training and Mobility of Researchers Programme (TMR). Ana Justel and Daniel

Peña acknowledge research support provided by DGES (Spain), under grants PB97–0021 and PB96–0111, respectively. The work of Ruey S. Tsay is supported by the National Science Foundation, the Chiang Chien-Kuo Foundation, and the Graduate School of Business, University of Chicago.

Appendix. Proof of Theorem 1

The conditional distribution of $\boldsymbol{\delta}_{j,k}$ given the sample and the other parameters is

$$P(\boldsymbol{\delta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\delta_{j,k}}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \cdot \alpha^{\boldsymbol{s}_{j,k}}(1-\alpha)^{k-\boldsymbol{s}_{j,k}}. \tag{A.1}$$

The likelihood function can be factorized as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) = f(\boldsymbol{y}_{p+1}^{j-1} \mid \boldsymbol{\theta}_{\delta_{j,k}}) \cdot f(\boldsymbol{y}_{j}^{T_{j,k}} \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \cdot f(\boldsymbol{y}_{T_{j,k}+1}^{n} \mid \boldsymbol{y}_{p+1}^{T_{j,k}}, \boldsymbol{\theta}_{\delta_{j,k}}),$$

where $\boldsymbol{y}_j^k = (y_j, \ldots, y_k)'$. Only $f(\boldsymbol{y}_j^{T_{j,k}} \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k})$ depends on $\boldsymbol{\delta}_{j,k}$ and it is the product of the conditional densities:

$$\begin{aligned}
f(y_j \mid \boldsymbol{y}_{p+1}^{j-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \delta_j) &\propto \exp\Big(-\frac{1}{2\sigma_a^2}(e_j(\mathbf{0}) - \delta_j\beta_j)^2\Big) \\
&\vdots \\
f(y_{j+k-1} \mid \boldsymbol{y}_{p+1}^{j+k-2}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) &\propto \exp\Big(-\frac{1}{2\sigma_a^2}(e_{j+k-1}(\mathbf{0}) - \delta_{j+k-1}\beta_{j+k-1} + \cdots \\
&\qquad +\pi_{k-1}\delta_j\beta_j)^2\Big) \\
f(y_{j+k} \mid \boldsymbol{y}_{p+1}^{j+k-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) &\propto \exp\Big(-\frac{1}{2\sigma_a^2}(e_{j+k}(\mathbf{0}) + \pi_1\delta_{j+k-1}\beta_{j+k-1} + \cdots \\
&\qquad +\pi_k\delta_j\beta_j)^2\Big) \\
&\vdots \\
f(y_{T_{j,k}} \mid \boldsymbol{y}_{p+1}^{T_{j,k}-1}, \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) &\propto \exp\Big(-\frac{1}{2\sigma_a^2}(e_{T_{j,k}}(\mathbf{0}) + \pi_{T_{j,k}-j-k+1}\delta_{j+k-1}\beta_{j+k-1} \\
&\qquad + \cdots + \pi_{T_{j,k}-j}\delta_j\beta_j)^2\Big).
\end{aligned}$$

Hence the likelihood function can be expressed as

$$\begin{aligned}
&f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (A.2)\\
&\propto \exp\Big(-\frac{1}{2\sigma_a^2}\Big(\sum_{t=j}^{j+k-1}(e_t(\mathbf{0}) + \sum_{i=0}^{t-j}\pi_i\delta_{t-i}\beta_{t-i})^2 + \sum_{t=j+k}^{T_{j,k}}(e_t(\mathbf{0}) + \sum_{i=t-j-k+1}^{t-j}\pi_i\delta_{t-i}\beta_{t-i})^2\Big)\Big),
\end{aligned}$$

and the residual $e_t(\boldsymbol{\delta}_{j,k})$ is given by

$$e_t(\boldsymbol{\delta}_{j,k}) = \begin{cases} e_t(\mathbf{0}) + \sum_{i=0}^{t-j} \pi_i \delta_{t-i} \beta_{t-i} & \text{if } t = j, \ldots, j+k-1 \\ e_t(\mathbf{0}) + \sum_{i=t-j-k+1}^{t-j} \pi_i \delta_{t-i} \beta_{t-i} & \text{if } t > j+k-1, \end{cases}$$

where $\pi_0 = -1$, $\pi_i = \phi_i$ for $i = 1, \ldots, p$ and $\pi_i = 0$ for $i < 0$ and $i > p$. Therefore, (A.2) can be written as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\delta_{j,k}}; \boldsymbol{\delta}_{j,k}) \propto \exp\Big(- \frac{1}{2\sigma_a^2} \sum_{t=j}^{T_{j,k}} e_t(\boldsymbol{\delta}_{j,k})^2 \Big).$$

Then by replacing in (A.1) we obtain (3.13) for any configuration of the vector $\boldsymbol{\delta}_{j,k}$.

The conditional distribution of $\boldsymbol{\beta}_{j,k}$ given the sample and the other parameters is

$$P(\boldsymbol{\beta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\beta_{j,k}}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\beta_{j,k}}; \boldsymbol{\beta}_{j,k}) \cdot P(\boldsymbol{\beta}_{j,k}).$$

Using (A.2)

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}_{\beta_{j,k}}; \boldsymbol{\beta}_{j,k}) \propto \exp\Big(- \frac{1}{2\sigma_a^2} \sum_{t=j}^{T_{j,k}} (e_t(\mathbf{0}) + \boldsymbol{\Pi}'_{t-j} \boldsymbol{D}_{j,k} \boldsymbol{\beta}_{j,k})' (e_t(\mathbf{0}) + \boldsymbol{\Pi}'_{t-j} \boldsymbol{D}_{j,k} \boldsymbol{\beta}_{j,k}) \Big).$$

Therefore,

$$\begin{aligned} & P(\boldsymbol{\beta}_{j,k} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\beta_{j,k}}) \\ & \propto \exp\Big(- \frac{1}{2\sigma_a^2} \sum_{t=j}^{T_{j,k}} (e_t(\mathbf{0}) + \boldsymbol{\Pi}'_{t-j} \boldsymbol{D}_{j,k} \boldsymbol{\beta}_{j,k})' (e_t(\mathbf{0}) + \boldsymbol{\Pi}'_{t-j} \boldsymbol{D}_{j,k} \boldsymbol{\beta}_{j,k}) \Big) \\ & \times \exp\Big(- \frac{1}{2\tau^2} (\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}_0)' (\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}_0) \Big) \\ & \propto \exp\Big(- \frac{1}{2} \Big(\boldsymbol{\beta}'_{j,k} \Big(\frac{1}{\sigma_a^2} \sum_{t=j}^{T_{j,k}} \boldsymbol{D}_{j,k} \boldsymbol{\Pi}_{t-j} \boldsymbol{\Pi}'_{t-j} \boldsymbol{D}_{j,k} + \frac{1}{\tau^2} \boldsymbol{I} \Big) \boldsymbol{\beta}_{j,k} \\ & \quad - 2 \Big(- \frac{1}{\sigma_a^2} \sum_{t=j}^{T_{j,k}} e_t(\mathbf{0}) \boldsymbol{\Pi}'_{t-j} \boldsymbol{D}_{j,k} + \frac{1}{\tau^2} \boldsymbol{\beta}'_0 \Big) \boldsymbol{\beta}_{j,k} \Big) \Big) \\ & \propto \exp\Big(- \frac{1}{2} (\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}^\star_{j,k})' \boldsymbol{\Omega}^{-1}_{j,k} (\boldsymbol{\beta}_{j,k} - \boldsymbol{\beta}^\star_{j,k}) \Big), \end{aligned}$$

where $\boldsymbol{\Omega}_{j,k}$ and $\boldsymbol{\beta}^\star_{j,k}$ are defined in (3.14) and (3.15) respectively.

References

Abraham, B. and Box, G. E. P. (1979). Bayesian analysis of some outlier problems in time series. *Biometrika* **66**, 229-236.

Barnett, G., Kohn, R. and Sheather, S. (1996). Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *J. Econom.* **74**, 237-254.

Barnett, G., Kohn, R. and Sheather, S. (1997). Robust Bayesian estimation of autoregressive-moving average models. *J. Time Ser. Anal.* **18**, 11-28.

Bruce, A. G. and Martin, D. (1989). Leave-k-out diagnostics for time series (with discussion). *J. Roy. Statist. Soc. Ser. B* **51**, 363–424.

Chang, I. and Tiao, G. C. (1983). Estimation of time series parameters in the presence of outliers. Technical Report 8, Statistics Research Center, University of Chicago.

Chang, I., Tiao, G. C. and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics* **3**, 193-204.

Chen, C. and Liu, L. (1993). Joint estimation of model parameters and outlier effects in time series. *J. Amer. Statist. Assoc.* **88**, 284-297.

Chow, G. C. (1960). A test for equality between sets of observations in two linear regressions. *Econometrica* **28**, 591-605.

Fox, A. J. (1972). Outliers in time series. *J. Roy. Statist. Soc. Ser. B* **34**, 350-363.

Hills, S. E. and Smith, A. F. M. (1992). Parameterization issues in Bayesian inference. In *Bayesian Statistics, 4* (Edited by J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), 641–649. Oxford University Press.

Justel, A. and Peña, D. (1996). Gibbs sampling will fail in outlier problems with strong masking. *J. Comput. Graph. Statist.* **5**, 176-189.

McCulloch, R. E. and Tsay, R. S. (1994). Bayesian analysis of autoregressive time series via the Gibbs sampler. *J. Time Ser. Anal.* **15**, 235-250.

Peña, D. (1987). Measuring the importance of outliers in ARIMA models. *New Perspectives in Theoretical and Applied Statistics* (Edited by Puri *et al.*), 109-118. John Wiley, New York.

Peña, D. (1990). Influential observations in time series. *J. Business Economic Statist.* **8**, 235-241.

Peña, D. and Maravall, A. (1991). Interpolation, outliers and inverse autocorrelations. *Comm. Statist.* (*Theory and Methods*) **20**, 3175-3186.

Sánchez, M. J. and Peña, D. (1997). The identification of multiple outliers in ARIMA models. Working Paper 97-76, Universidad Carlos III de Madrid.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. *J. Amer. Statist. Assoc.* **81**, 132-141.

Tsay, R. S. (1988). Outliers, level shifts, and variance change in time series. *J. Forecasting* **7**, 1-20.

Tsay, R. S., Peña, D. and Pankratz, A. E. (2000). Outliers in multivariate time series. *Biometrika*, **87**, 789-804.

Department of Mathematics, Universidad Autónoma de Madrid, Spain.

E-mail: ana.justel@uam.es

Department of Statistics and Econometrics, Universidad Carlos III de Madrid, Spain.

E-mail: dpena@est-econ.uc3m.es

Graduate School of Business, University of Chicago, Spain.

E-mail: rst@gsbrst.uchicago.edu

(Received September 1999; accepted October 2000)