Journal of the Korean Statistical Society (1999), 28: 3, pp 311–324

Bayesian Outlier Detection in Regression Model

Younshik Chung[1] and Hyungsoon Kim[2]

ABSTRACT

The problem of 'outliers', observations which look suspicious in some way, has long been one of the most concern in the statistical structure to experimenters and data analysts. We propose a model for an outlier problem and also analyze it in linear regression model using a Bayesian approach. Then we use the mean-shift model and SSVS(George and McCulloch, 1993)'s idea which is based on the data augmentation method. The advantage of proposed method is to find a subset of data which is most suspicious in the given model by the posterior probability. The MCMC method(Gibbs sampler) can be used to overcome the complicated Bayesian computation. Finally, a proposed method is applied to a simulated data and a real data.

Keywords: Gibbs sampler; Latent variable; Linear mixed normal model; Linear regression model; Mean-shift model; Outlier; Variance-inflation model.

1. INTRODUCTION

The problem of 'outliers', observations which look suspicious in some way, has long been one of the most concern in the statistical structure to experimenters and data analysts. In this paper, we propose a model for an outlier problem and also analyze it using a Bayesian approach. The Bayesian approaches for outlier detection can be classified to two procedures such as using alternative model for outliers or not. For the method without having alternative model, the predictive distribution is used in Geisser(1985) and Pettit and Smith(1985), or the posterior distribution used in Johnson and Geisser(1983), Chaloner and Brant(1988) and Guttman and Pena(1993). For alternative model, the mean-shift model or the variance-inflation model is used in Guttman(1973) and Sharples(1990). Let Y be an observation vector from $N(\mu, \sigma^2)$. The mean-shift model is that a suspicious observation is distributed as $N(\mu + m, \sigma^2)$. If m is not a zero, the corresponding observation is decided as an outlier, Guttman(1973) applied the mean-shift

[1]Department of Statistics, Pusan National University, Pusan, 609-735 Korea
[2]Department of Statistics, Pusan National University, Pusan, 609-735 Korea

model to a linear model. The variance-inflation model is that an observation y_i be from $N(\mu, b_i\sigma^2)$ where the observation, y_i, with $b_i >> 1$, is treated as an outlier (Box and Tiao, 1968). Sharples(1990) showed how variance inflation can be incorporated easily into general hierarchical models, retaining tractability of analysis.

In this paper, we use the mean-shift model in linear regression model. Specifically, we assume that for some particular $(n \times p)$ matrix X of constants(or design matrix), it is intended to generate data $Y = (Y_1, \cdots, Y_n)^t$, such that

$$Y = X\beta + \epsilon \tag{1.1}$$

where $\beta = (\beta_1, \cdots, \beta_p)^t$ is a set of p unknown regression parameters, and where the $(n \times 1)$ error vector ϵ is normally distributed with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2 I_n$, where σ^2 is unknown. Despite the fact that (1.1) is intended generation scheme, it is feared that departures from this model will occur, indeed that the observations y may be generated as follows; for $i = 1, \cdots, n$,

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + m_i + \epsilon_i. \tag{1.2}$$

Therefore if $m_i = 0$, then ith observation is not an outlier. Otherwise, it is considered as an outlier. For detecting outliers, we use SSVS(stochastic search variable selection) method in George and McCulloch(1993). The SSVS was introduced as the method for selecting the best predictors in multiple regression model using Gibbs sampler. In general likelihood approaches, we find one outlier which is most suspicious and next the pair and the triple of outliers, and so on. Then we can not compare the one most outlier with a pair of most outliers and there exits sometimes masking. But our Bayesian method overcomes such problems. In this procedure, the most great advantage is that we can find the set of outliers which is most suspecious among the data.

The plan of this article is as follows. In section 2, in order to find the outliers, we introduce and motivate the hierarchical framework that depends on the SSVS in George and McCulloch(1993). In section 3, we illustrate our proposed method on a simulated example and a real data set (Darwin's data: Box and Tiao, 1973). Finally, in Section 4, we extend our proposed method to normal linear mixed model.

2. HIERARCHICAL BAYESIAN FORMULATION

2.1. SSVS For Outliers

For detecting outliers in linear regression model, we apply the mean-shift model to linear regression model. Despite the fact that (1.1) is intended generation scheme, it is feared that departures from this model will occur, indeed that the observations y may be generated as follows; for $i = 1, \cdots, n$,

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + m_i + \epsilon_i. \tag{2.1}$$

Therefore our model is reexpressed as follows;

$$Y = \tilde{X}\tilde{\beta} + \epsilon \tag{2.2}$$

where $\tilde{X} = [X, I_n]$ and $\tilde{\beta} = (\beta_1, \cdots, \beta_p, m_1, \cdots, m_n)^t$ and I_n denotes the $n \times n$ identity matrix.

By introducing the latent variable $\gamma_j = 0$ or 1, we represent our normal mixture by

$$m_j \mid \gamma_j \sim (1-\gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2) \tag{2.3}$$

and

$$Pr(\gamma_j = 1) = 1 - Pr(\gamma_j = 0) = p_j. \tag{2.4}$$

This is based on the data augmentation idea of Tanner and Wong(1987). George and McCulloch(1993) use the same structure (2.3) and (2.4) for selecting variables in linear regression model. Diebolt and Robert(1994) have also successfully used this approach for selecting the number of components in the mixture distribution.

When $\gamma_i = 0$, $m_i \sim N(0, \tau_i^2)$, and when $\gamma_i = 1$, $m_i \sim N(0, c_i^2\tau_i^2)$. Following George and McCulloch(1993), first, we set $\tau_i(> 0)$small so that if $\gamma_i = 0$, then m_i would probably be so small that it could be "safely" estimated by 0. Second, we set c_i large ($c_i > 1$ always) so that if $\gamma_i = 1$ then a non-zero estimate of m_i means that the corresponding data y_i should probably be an outlier in the given model.

To obtain (2.3) as the prior for $m_i \mid \gamma_i$, we use a multivariate normal prior

$$m \mid \gamma \sim N_k(0, D_\gamma R D_\gamma), \tag{2.5}$$

where $m = (m_1, \cdots, m_n)$, $\gamma = (\gamma_1, \cdots, \gamma_k)$, R is the prior correlation matrix, and

$$D_\gamma = diag[a_1\tau_1, \ldots, a_k\tau_k] \tag{2.6}$$

with $a_i = 1$ if $\gamma_i = 0$ and $a_i = c_i$ if $\gamma_i = 1$. For selecting the values of tunning factors c_1 and τ_i, see George and McCulloch(1993) and section 2.2. Also see George and McCulloch(1993) for selection of R. As a particular interest, the identity matrix can be used for R. The choice of $f(\gamma)$ should incorporate any available prior information about which subsets of $y_1, \cdots, y_n$ should be outliers in the given model. Although this may seem difficult with 2^n possible choices, especially with large n, symmetry considerations may simplify this work. For example, a reasonable choice might have the γ's independent with marginal distributions (2.4), so that

$$f(\gamma \mid p_1, \cdots, p_n) = \prod_{i=1}^{n} p_i^{\gamma_i}(1 - p_i)^{(1-\gamma_i)}. \tag{2.7}$$

Although (2.7) implies that the outlier of y_i is independent on the outlier of y_j for all $i \neq j$, we found it to work well in the various situations. The uniform or indifference prior $f(\gamma) = 2^{-n}$ is the special case of (2.7) where each y_i has an equal chance to be an outlier.

Also, it is assumed that the prior of $\beta = (\beta_1, \cdots, \beta_p)$ is normal distribution with mean vector $\mu = (\mu_1, \cdots, \mu_p)$ and variance-covariance matrix Σ^{-1}. That is,

$$\beta \sim N(\mu, \Sigma^{-1}). \tag{2.8}$$

Finally, we use the inverse gamma conjugate prior

$$\sigma^2 \mid \gamma \sim IG(\frac{\nu_\gamma}{2}, \frac{\nu_\gamma \lambda_\gamma}{2}). \tag{2.9}$$

In this procedure, our main concern for embedding the normal linear model (1.2) into the hierarchical mixture model is to obtain the marginal posterior distribution $f(\gamma \mid Y) \propto f(Y \mid \gamma)f(\gamma)$, which contains the information relevant to outliers. As mentioned as before, $f(\gamma)$ may be interpretated as the statistician's prior probability that the y_i's corresponding to an non-zero components of γ should be outliers in the given model. The posterior density $f(\gamma \mid Y)$ updates the prior probabilities on each of the 2^n possible values of γ. Identifying each γ with a subset of data via that $\gamma_i = 1$ is equivalent to that y_i is an outlier, those γ with higer posterior probability $f(\gamma \mid Y)$ identify the subset of data which is most suspicious by data and the statistician's prior information. Therefore, $f(\gamma \mid Y)$ provides a ranking that can be used to select a subset of the most suspicious data.

2.2. Choice of c_j and τ_j

As an idea discussed in George and McCulloch(1993), the choice of c_j and τ_j should be based on two considerations. First, the choice determines a neighborhood of 0 where SSVS treats m_j as equivalent to zero. This neighborhood is obtained as $(-\delta_j, \delta_j)$, where $-\delta_j$ and δ_j are the intersection points of the densities $N(0, \tau_j^2)$ and $N(0, c_j^2\tau_j^2)$. Because $(-\delta_j, \delta_j)$ is the interval where $N(0, \tau_j^2)$ dominates $N(0, c_j^2\tau_j^2)$, the posterior probability of $\gamma_j = 0$ is more likely to be large when m_j is in $(-\delta_j, \delta_j)$. Thus SSVS entails estimating m_j by 0, according to whether m_j is in $(-\delta_j, \delta_j)$. Second, the choice determines how different the two densities are. If $N(0, \tau_j^2)$ is too peaked or $N(0, c_j^2\tau_j^2)$ is too spread out, the Gibbs sampler may converge too slowly when the data are not very informative.

Our recommentation is to first choose δ_j. This may be obtained by practical considerations such as setting δ_j equal to the largest value of $\mid m_j \mid$ for which a plausible change in y_j would make no practical difference. Once δ_j has been chosen, we recommand choosing c_j between 10 and 100. This range for c_j seems to provide seperation between $N(0, \tau_j^2)$ and $N(0, c_j^2\tau_j^2)$ which is large enough to yield a useful posterior and small enough to avoid Gibbs sampling convergence problems. When such prior information is available, we also recommand choosing c_j so that $N(0, c_j^2\tau_j^2)$ give reasonable probability to all plausible values of m_j. Finally, τ_j is obtained from δ_j and c_j by $\tau_j = [2log(c_j)c_j^2/(c_j - 1)]^{-\frac{1}{2}}\delta_j$. Note that for $c_j = 10, \delta_j = 2.15\tau_j$ and for $c_j = 100, \delta_j = 3.04\tau_j$. For selecting the regression parameters in linear model, a similar semiautomatic strategy for selecting τ_j and c_j based on statistical significance is described in George and McCulloch(1993).

2.3. Full Conditional Distributions

Densities are denoted generically by brackets, so joint, conditional, and marginal forms, for example, appear as $[X, Y], [X \mid Y]$ and $[X]$, respectively. The joint posterior density of $\beta, m, \gamma, \sigma^2$ given $y_1, \cdots, y_n$ is given by

$$
\begin{aligned}
[\beta, m, \gamma, \sigma^2 \mid y_1, \cdots, y_n] \;\; \propto \;\; & \frac{1}{(2\pi)^{\frac{n}{2}} \mid \sigma^2 I \mid^{\frac{1}{2}}} \exp\{-\frac{1}{2}(Y - \tilde{X}\tilde{\beta})^t \sigma^{-2}(Y - \tilde{X}\tilde{\beta})\} \\
\times \;\; & \frac{1}{(2\pi) \mid \Sigma^{-1} \mid^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\beta - \mu)^t \Sigma(\beta - \mu)\} \\
\times \;\; & \frac{1}{(2\pi)^{\frac{n}{2}} \mid D_\gamma R D_\gamma \mid^{\frac{1}{2}}} \exp\{-\frac{1}{2} m^t (D_\gamma R D_\gamma)^{-1} m\}
\end{aligned}
$$

$$\times \quad \frac{(V_\gamma\lambda_\gamma/2)^{\frac{\lambda_\gamma}{2}}}{\Gamma(V_\gamma/2)} \frac{1}{(\sigma^2)^{\frac{V_\gamma}{2}+1}} \exp\{-\frac{V_\gamma\lambda_\gamma}{2\sigma^2}\}$$
$$\times \quad \prod_{i=1}^{n} p_i^{\gamma_i}(1-p_i)^{1-\gamma_i}. \tag{2.10}$$

In order to Gibbs sampler, the full conditional distributions are needed as follows;

$$[\beta \mid m,\gamma,\sigma^2,y_1,\cdots,y_n] \propto \exp\{-\frac{1}{2}(\sigma^{-2}\tilde{\beta}^t\tilde{X}^t\tilde{X}\tilde{\beta} - 2\sigma^{-2}\tilde{\beta}^t\tilde{X}^tY + \beta^t\Sigma\beta - 2\beta^t\Sigma\mu)\}, \tag{2.11}$$

$$[m \mid \beta,\gamma,\sigma^2,y_1,\cdots,y_n] \propto \exp\{-\frac{1}{2}(\sigma^{-2}\tilde{\beta}^t\tilde{X}^t\tilde{X}\tilde{\beta} - 2\sigma^{-2}\tilde{\beta}^t\tilde{X}^tY + m^t(D_\gamma R D_\gamma)^{-1}m)\}, \tag{2.12}$$

and

$$[\sigma^2 \mid \beta,m,\gamma,y_1,\cdots,y_n] = IG(\frac{n+V_\gamma}{2}, \frac{|Y-\tilde{X}\tilde{\beta}|^2 + V_\gamma\lambda_\gamma}{2}). \tag{2.13}$$

Finally, the vector γ is obtained componentwise by sampling consecutively from the conditional distribution

$$\gamma_i \sim [\gamma_i \mid y_1,\cdots,y_n,\beta,m,\sigma^2,\gamma_{(-i)}] = [\gamma_i \mid m,\sigma^2,\gamma_{(-i)}] \tag{2.14}$$

where $\gamma_{(-i)} = (\gamma_1,\cdots,\gamma_{i-1},\gamma_{i+1},\cdots,\gamma_n)$. The last equality in (2.14) holds because the independency of (2.14) on Y results from the hierarchical structure where γ affects Y only through m and β is independent of γ by the assumption above. Since $[m,\gamma_i,\gamma_{(-i)},\sigma^2] = [m \mid \gamma_i,\gamma_{(-i)},\sigma^2][\sigma^2 \mid \gamma_i,\gamma_{(-i)}][\gamma_i,\gamma_{(-i)}]$,

$$\begin{aligned}[\gamma_i = 1 \mid y_1,\cdots,y_n,\beta,m,\sigma^2,\gamma_{(-i)}] &= \frac{[m,\gamma_i=1,\gamma_{(-i)},\sigma^2]}{\sum_{k=0}^{1}[m,\gamma_i=k,\gamma_{(-i)},\sigma^2]} \\ &= \frac{a}{a+b}\end{aligned} \tag{2.15}$$

where $a = [m \mid \gamma_i = 1,\gamma_{(-i)},\sigma^2][\sigma^2 \mid \gamma_i = 1,\gamma_{(-i)}][\gamma_i = 1,\gamma_{(-i)}]$and $b = [m \mid \gamma_i = 0,\gamma_{(-i)},\sigma^2][\sigma^2 \mid \gamma_i = 0,\gamma_{(-i)}][\gamma_i = 0,\gamma_{(-i)}]$. Note that under the prior (2.7) on γ and when the prior parameters for σ^2 in (2.9) are constant, then (3.6) can be obtained more simply by

$$a = [m \mid \gamma_i = 1,\gamma_{(-i)}]p_i, \qquad b = [m \mid \gamma_i = 0,\gamma_{(-i)}](1-p_i). \tag{2.16}$$

3. ILLUSTRATIVE EXAMPLES

3.1. Simulate data

In this section we illustrate the performance of SSVS on simulated examples. We consider the simple linear regression model, that is,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \cdots, n \tag{3.1}$$

with $n = 10$ and let $m_i = 0$ for $i \neq 7$ and $m_7 = 10$. $x_i \sim N(0,1)$, $\epsilon_i \sim N(0,1)$ for $i = 1, \cdots, 10$ and $(\beta_0, \beta_1) = (0.5, 1)$. Therefore since $m_7 = 10$, we assume that observation 7 is an outlier in this data set. For the notational convenience, let

$$\mu_{\beta_0} = \frac{\sum y_i + \sigma^2(\mu_0\sigma_{11} + \mu_1\sigma_{12}) - \beta_1(\sum x_i + \sigma^2\sigma_{12}) - \sum m_i}{n + \sigma^2\sigma_{11}}$$

and

$$\mu_{\beta_1} = \frac{\sum x_i y_i + \sigma^2(\mu_0\sigma_{21} + \mu_1\sigma_{22}) - \beta_0(\sum x_i + \sigma^2\sigma_{12}) - \sum m_i x_i}{\sum x_i + \sigma^2\sigma_{22}}$$

Then

$$[\beta_0 \mid \beta_1, m, \gamma, \sigma^2, y_1, \cdots, y_n] = N(\mu_{\beta_0}, \sigma^2) \tag{3.2}$$

and

$$[\beta_1 \mid \beta_0, m, \gamma, \sigma^2, y_1, \cdots, y_n] = N(\mu_{\beta_1}, \sigma^2). \tag{3.3}$$

For $i = 1, \cdots, n$,

$$[m_i \mid \beta_0, \beta_1, m_{(-i)}, \gamma, \sigma^2, y_1, \cdots, y_n] = N(\mu_{m_i}, \frac{1}{\sigma^{-2} + (a_i\tau_i)^{-2}}) \tag{3.4}$$

where $\mu_{m_i} = \frac{y_i - \beta_0 - \beta_1 x_i}{1 + \sigma^2(a_i\tau_i)^{-2}}$.

$$[\sigma^2 \mid \beta, m, \gamma, y_1, \cdots, y_n] = IG(\frac{n}{2}, \frac{|Y - \tilde{X}\tilde{\beta}|^2}{2}) \tag{3.5}$$

and

$$[\gamma_i = 1 \mid y_1, \cdots, y_n, \beta, m, \sigma^2, \gamma_{(-i)}] = \frac{a}{a+b} \tag{3.6}$$

where $a = [m \mid \gamma_i = 1, \gamma_{(-i)}]p_i, b = [m \mid \gamma_i = 0, \gamma_{(-i)}](1 - p_i)$.

Table 3.1:

outlier numbers	observation	proportion
0		0.26
	2	0.01
	3	0.01
	4	0.01
1	6	0.01
	7	0.36
	10	0.05
	1, 7	0.01
	2, 7	0.01
	4, 7	0.01
2	5, 7	0.01
	7, 9	0.01
	7, 10	0.03
	9, 10	0.01
3	1, 2, 5	0.01
	1, 6, 7	0.01
4	1, 4, 6, 10	0.01
	1, 7, 8, 9	0.01

We applied SSVS to our model with the indifference prior $f(\gamma) = (\frac{1}{2})^{10}$. $\tau_1 = \cdots = \tau_{10} = 0.25, c_1 = \cdots = c_{10} = 10$, $R = I$, and $\nu_\gamma = 0$. In Table 3.1, observation 7 is considered as outlier since its correponding posterior probability is 0.36 which is highest among all data. Also, we may consider that there is no outliers in data set since its posterior probability is second highest.

3.2. Darwin's Data

Consider the analysis of Darwin's data on the difference in heights of self- and cross-fertilized plants quoted by Box and Tiao(1973, p153). The data consists of measurements on 15 pairs of plants. Each pair contained a self-fertilized and cross-fertilized plant grown in the same pot and from the same seed. Arranged for convenience in order of magnitude, the n=15 observations (on differences in in heights in eighths of an inch of self-fertilized and cross-fertilized plants) are:

-67, -48, 6, 8, 14, 16, 23, 24, 28, 29, 41, 49, 56, 60, 75. Guttman, Dutter and Freeman(1978) re-examine the Darwin's data to detect outlier(s) using Bayesian approach with the model as follows;

$$Y = \beta \mathbf{1} + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n) \tag{3.7}$$

where $\mathbf{1} = (1, \ldots, 1)^t$. Guttman et al(1978) mentioned that observations 1 and 2, having values -67 and -48, are identified as spurious observations since they have the highest posterior probability. Table 3.2 shows that observations 1 and 2, having values -67 and -48 respectively, are considered as outliers since they have the highest posterior probability 0.57. Also, observations 1, 2 and 7 may be considered as outliers.

Table 3.2:

outlier numbers	observation	proportion
0		0.00
1		0.00
2	1, 2	0.57
3	1, 2, 7	0.38
	1, 2, 8	0.01
4	1, 2, 7, 8	0.04

4. EXTENSION TO LINEAR MIXED NORMAL MODEL

In this section, we use the mean-shift model in linear mixed model. Specifically, we assume that for some particular $(n \times p)$ matrix X and $(n \times l)$ matrix Z of constants(or design matrices), it is intended to generate data $Y = (Y_1, \cdots, Y_n)^t$, such that

$$Y = X\beta + Zb + \epsilon \tag{4.1}$$

where $\beta = (\beta_1, \cdots, \beta_p)^t$ and $b = (b_1, \ldots, b_l)$. And we assume that the $(n \times 1)$ error vector ϵ is normally distributed with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2 I_n$, where σ^2 is unknown and the $(l \times 1)$ vector b is assumed to be normal with zero mean vector and the variance-covariance matrix $\sigma_b^2 I_l$. Also the independency of b and ϵ are assumed. For detecting the outliers, we consider that observations

y may be generated as follows; for $i = 1, \cdots, n$,

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j=1}^{l} z_{ij}b_j + m_i + \epsilon_i. \tag{4.2}$$

Therefore if $m_i = 0$, then ith observation is not an utlier. Otherwise, it is considered as an outlier. For detecting outliers, we use SSVS(stochstic search variable selection) method in section 2.

Therefore our model is reexpressed as follows;

$$Y = \tilde{X}\tilde{\beta} + Zb + \epsilon \tag{4.3}$$

where $\tilde{X} = [X, I_n]$ and $\tilde{\beta} = (\beta_1, \cdots, \beta_p, m_1, \cdots, m_n)^t$ and I_n denotes the $n \times n$ identity matrix.

As before, we use all forms in (2.3) - (2.7) into our model in (4.3). Also, it is assumed that the prior of $\beta = (\beta_1, \cdots, \beta_p)$ is normal distribution with mean vector $\mu = (\mu_1, \cdots, \mu_p)$ and variance-covariance matrix Σ^{-1}. That is,

$$\beta \sim N(\mu, \Sigma^{-1}). \tag{4.4}$$

Finally, we use the inverse gamma conjugate prior

$$\sigma^2 \mid \gamma \sim IG(\frac{\nu_\gamma}{2}, \frac{\nu_\gamma\lambda_\gamma}{2}), \quad \sigma_b^2 \sim IG(\frac{\nu_b}{2}, \frac{\nu_b\lambda_b}{2}). \tag{4.5}$$

4.1. Full Conditional Densities

The joint posterior density of $\beta, b, m, \gamma, \sigma^2$ given $y_1, \cdots, y_n$ is given by

$$\begin{aligned}
[\beta, b, m, \gamma, \sigma^2, \sigma_b^2 \quad &| \quad y_1, \cdots, y_n] \\
&\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\{-\frac{1}{2\sigma^2}(Y - \tilde{X}\tilde{\beta} - Zb)^t(Y - \tilde{X}\tilde{\beta} - Zb)\} \\
&\times \frac{1}{\mid \Sigma^{-1} \mid^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\beta - \mu)^t\Sigma(\beta - \mu)\} \\
&\times \frac{1}{(2\pi)^{\frac{n}{2}} \mid D_\gamma R D_\gamma \mid^{\frac{1}{2}}} \exp\{-\frac{1}{2}m^t(D_\gamma R D_\gamma)^{-1}m\} \\
&\times \frac{1}{(\sigma^2)^{\frac{V}{2}+1}} \exp\{-\frac{V\lambda}{2\sigma^2}\} \times \frac{1}{(\sigma_b^2)^{\frac{V_b}{2}+1}} \exp\{-\frac{V_b\lambda_b}{2\sigma_b^2}\} \\
&\times \prod_{i=1}^{n} p_i^{\gamma_i}(1 - p_i)^{1-\gamma_i} \times (\sigma_b^2)^{-\frac{l}{2}} \exp\{-\frac{b^t b}{2\sigma_b^2}\}.
\end{aligned} \tag{4.6}$$

In order to Gibbs sampler, the full conditional distributions are needed as follows;

$$\begin{aligned}[\beta \quad &| \quad b, m, \gamma, \sigma^2, y_1, \cdots, y_n] \\ &\propto \quad \exp\{-\frac{1}{2\sigma^2}(\tilde{\beta}^t\tilde{X}^t\tilde{X}\tilde{\beta} - 2\tilde{\beta}^t\tilde{X}^tY + 2\tilde{\beta}^t\tilde{X}^tZb) - \frac{1}{2}(\beta^t\Sigma\beta - 2\beta^t\Sigma\mu)\},\end{aligned} \tag{4.7}$$

$$\begin{aligned}[b \quad &| \quad \beta, m, \gamma, \sigma^2, \sigma_b^2, y_1, \cdots, y_n] \\ &\propto \quad \exp\{-\frac{1}{2\sigma^2\sigma_b^2}[\sigma_b^2(-2Y^tZb + 2\tilde{\beta}^t\tilde{X}Zb + b^tZ^tZb) + \sigma^2b^tb]\},\end{aligned} \tag{4.8}$$

$$\begin{aligned}[m \quad &| \quad \beta, b, \gamma, \sigma^2, \sigma_b^2, y_1, \cdots, y_n] \\ &\propto \quad \exp\{-\frac{1}{2\sigma^2}(Y - \tilde{X}\tilde{\beta} - Zb)^t(Y - \tilde{X}\tilde{\beta} - Zb) - \frac{1}{2}m^t(D_\gamma RD_\gamma)^{-1}m\},\end{aligned} \tag{4.9}$$

$$[\sigma^2 \mid \beta, b, m, \sigma_b^2, \gamma, y_1, \cdots, y_n] = IG(\frac{l + V_b}{2}, \frac{|Y - \tilde{X}\tilde{\beta} - Zb|^2 + V_\gamma\lambda_\gamma}{2}), \tag{4.10}$$

and

$$[\sigma_b^2 \mid \beta, b, m, \sigma^2, \gamma, y_1, \cdots, y_n] = IG(\frac{n + V_\gamma}{2}, \frac{b^tb + V_b\lambda_b}{2}). \tag{4.11}$$

Finally, the vector γ is defined in (2.15) and (2.16).

4.2. Generated data

In this section we illustrate the performance of SSVS on simulated example. We consider the simple linear mixed normal model, that is,

$$y_i = \beta_0 + \beta_1 x_i + b_1 z_{i1} + b_2 z_{i2} + \epsilon_i, \qquad i = 1, \cdots, n \tag{4.12}$$

with $n = 10$, $x_i, z_{i1}, z_{i2} \sim N(0,1)$, $\epsilon_i \sim N(0,1)$ for $i = 1, \cdots, 10$ and $(\beta_0, \beta_1) = (0.5, 1)$ and $b_i \sim N(0,1)$ for $i = 1, 2$. And let $m_7 = 10$ and $m_i = 0$ for $i \neq 7$. Therefore we assume that observation 7 is outlier in this data set. For the notational convenience, let

$$\mu_{\beta_0} = \frac{\sum y_i + \sigma^2(\mu_0\sigma_{11} + \mu_1\sigma_{12}) - \beta_1(\sum x_i + \sigma^2\sigma_{12}) - b_1\sum z_{i1} - b_2\sum z_{i2} - \sum m_i}{n + \sigma^2\sigma_{11}},$$

$$\mu_{\beta_1} = \frac{\sigma^2(\mu_0\sigma_{21} + \mu_1\sigma_{22}) - \beta_0(\sum x_i + \sigma^2\sigma_{12}) - \sum(b_1x_iz_{i1} + b_2x_iz_{i2} + m_ix_i - x_iy_i)}{\sum x_i^2 + \sigma^2\sigma_{22}},$$

$$\mu_{b_1} = \frac{\sigma_b^2 \sum [y_i z_{i1} - (\beta_0 + \beta_1 x_i + m_i) z_{i1} - b_2 z_{i1} z_{i2}]}{\sigma_b^2 \sum z_{i1}^2 + \sigma^2},$$

and

$$\mu_{b_2} = \frac{\sigma_b^2 \sum [y_i z_{i2} - (\beta_0 + \beta_1 x_i + m_i) z_{i2} - b_1 z_{i1} z_{i2}]}{\sigma_b^2 \sum z_{i2}^2 + \sigma^2}.$$

Then

$$[\beta_0 \mid \beta_1, b, m, \gamma, \sigma^2, y_1, \cdots, y_n] = N(\mu_{\beta_0}, \frac{\sigma^2}{n + \sigma^2 \sigma_{11}}), \tag{4.13}$$

$$[\beta_1 \mid \beta_0, b, m, \gamma, \sigma^2, y_1, \cdots, y_n] = N(\mu_{\beta_1}, \frac{\sigma^2}{\sum x_i^2 + \sigma^2 \sigma_{22}}), \tag{4.14}$$

$$[b_1 \mid b_2, \beta, m, \gamma, \sigma^2, y_1, \cdots, y_n] = N(\mu_{b_1}, \frac{\sigma^2 \sigma_b^2}{\sigma_b^2 \sum z_{i1}^2 + \sigma^2}), \tag{4.15}$$

and

$$[b_2 \mid b_1, \beta, m, \gamma, \sigma^2, y_1, \cdots, y_n] = N(\mu_{b_2}, \frac{\sigma^2 \sigma_b^2}{\sigma_b^2 \sum z_{i2}^2 + \sigma^2}). \tag{4.16}$$

For $i = 1, \cdots, n$,

$$[m_i \mid \beta_0, \beta_1, m_{(-i)}, \gamma, \sigma^2, y_1, \cdots, y_n] = N(\mu_{m_i}, \frac{1}{\sigma^{-2} + (a_i \tau_i)^{-2}}) \tag{4.17}$$

where $\mu_{m_i} = \frac{y_i - \beta_0 - \beta_1 x_i}{1 + \sigma^2 (a_i \tau_i)^{-2}}$.

$$[\sigma^2 \mid \beta, m, \gamma, y_1, \cdots, y_n] = IG(\frac{n}{2}, \frac{|Y - \tilde{X}\tilde{\beta}|^2}{2}) \tag{4.18}$$

and

$$[\gamma_i = 1 \mid y_1, \cdots, y_n, \beta, m, \sigma^2, \gamma_{(-i)}] = \frac{a}{a+b} \tag{4.19}$$

where $a = [m \mid \gamma_i = 1, \gamma_{(-i)}] p_i, b = [m \mid \gamma_i = 0, \gamma_{(-i)}](1 - p_i)$. We applied SSVS to our model with the indifference prior $f(\gamma) = (\frac{1}{2})^{10}. \tau_1 = \cdots = \tau_{10} = 0.25, c_1 = \cdots = c_{10} = 10$, $R = I$, and $\nu_\gamma = 0$ and $\nu_b = 0$. Table 4.1 shows that observation 7 is considered as an outlier since its proportion (which is an approximate posterior probability) is highest. Also, it may be assumed that there are no outliers.

Table 4.1:

outlier numbers	observation	proportion
0		0.23
	2	0.09
	4	0.03
	6	0.01
1	7	0.35
	8	0.07
	10	0.01
	2, 4	0.01
	2, 7	0.06
	2, 9	0.01
2	4, 6	0.02
	4, 8	0.02
	6, 7	0.01
	6, 10	0.02
3	2, 4, 8	0.01
	4, 8, 7	0.01
4	2, 4, 9, 10	0.01
	1, 7, 8, 9	0.01

References

Box, G. E. P. and Tiao, G. C. (1968), "A Bayesian Approach to Some Outlier Problems," *Biometrika,* **55,** 119-129.

Box, G. E. P. and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis,* Addison-Wesley Publishing Co., U.S.A.

Chaloner, K. and Brant, R. (1988), "A Bayesian Approach to Outlier Detection and Residual Analysis," *Biometrika,* **75,** 651-659.

Diebolt, J. and Robert, C. (1994), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of royal Statistical Society,* Ser. B, 363-375.

Geisser, S. (1985), "On the Predicting of Observables: A Selective Update," in *Bayesian Statistics 2*, Ed. Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M., 203-230, Amsterdam: North Holland.

Gelfand, A. and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* **85,** 398-409.

George, E.I. and McCulloch, R.E. (1993), "Variable Selection Via Gibbs Sampling," *Journal of American Statistical Association*, **88**, 881-889.

Guttman, I. (1973), "Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity - A Bayesian Approach," *Technometrics,* **15,** 4, 723-738.

Guttman, I., Dutter, R. and Freeman, P. R. (1978), "Care and Handling of Univariate Outliers in the General Linear Model to Detect Spuriosity - A Bayesian Approach," *Technometrics,* **20,** 2, 187-193.

Guttman, I. and Pena, D. (1993), "A Bayesian Look at Diagnostics in the Univariate Linear Model," *Statistical Sinica,* **3**, 367-390.

Johnson, W. and Geisser, S. (1983), "A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis," *Journal of the American Statistical Association,* **78,** 137-144.

Pettit, L. I. and Smith, A. F. M. (1985), "Outliers and Influential Observations in Linear Models," in *Bayesian Statistics 2,* Ed. Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M., 473-494, Amsterdam: North Holland.

Sharples, L. D. (1990) "Identification and Accommodation of Outliers in General Hierarchical Models," *Biometrika,* **77,** 3, 445-453.

Tanner, M. and Wong, W. (1987), "The Calculation of Posterior Distributions by Data Augmentation (with discussion)," *Journal of the American Statistical Association,* **82,** 528-550.