Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(2) 2010 243-253

A Simple Method for Bayesian Robust Estimation

Guan Xing[1] and J. Sunil Rao[2*]

[1]*Bristol-Myers Squibb, New York, USA*

[2]*University of Miami, Miami, USA*

Received 13 April 2010; Revised 14 May 2010; Accepted 14 May 2010

SUMMARY

We introduce a new Bayesian robust estimation approach to deal with contaminated data. The formulation is based on latent indicator variables which are used to down-weight potential outliers. The posterior distributions (and functionals) of the parameters of interest and the indicator variables are derived using a Gibbs sampler. A diagnostic plot from the posterior distribution of the latent variables provides visual evidence of the relative weights attached to each observation. This approach is simple and rather general in its applicability. We show examples from linear and generalized linear regression, as well as multivariate estimation.

Keywords : Bayesian, Robust estimation, Gibbs sampler.

1. INTRODUCTION

When not all observations follow the assumed idealized distribution, the data are termed contaminated. Classical maximum likelihood estimation does not work well under this situation, in the sense that the outliers can have substantial influence on the estimates. More robust distributions are suggested to fit the model, such as replacing a normal distribution with a t distribution for heavy tailed data, or using a negative binomial distribution instead of a poisson distribution for the over-dispersed counts. Another strategy is to develop special robust estimation methods to deal with this problem. A common contamination model is

$$(1 - \alpha) * f_1(x) + \alpha * f_2(x) \qquad (1)$$

which assumes most observations come from distribution $f_1(x)$, while a few "bad" observations come from distribution $f_2(x)$. The probability of an observation being "bad" is taken to be α.

From a Bayesian viewpoint, Box and Tiao (1968) and Justel and Peña (1996, 2001) suggested the variance-inflation model $(1 - \alpha) * N(\theta, \sigma^2) + \alpha * N(\theta, k^2\sigma^2)$; Guttman (1973) and Verdinelli and Wasserman (1991) suggested the mean-shift model $(1 - \alpha) * N(\theta, \sigma^2) + \alpha * N(\theta + \lambda, \sigma^2)$, where k and λ are positive constants. Both models are special cases of the common contamination model (1). However, these robust models assume that possible outliers come from a normal distribution with different location parameter or large variance, which may not be a good approximation to the true situation sometimes. In addition, the choice for the location shift constant or the variance-inflation constant could affect the final inference.

To extend these Bayesian robust models to more general cases, we propose a new approach using latent Bernoulli indicator variable attached to each observation. If the indicator variable takes the value 1, the corresponding observation will be included in the model construction. Otherwise, the corresponding observation will not participate in the modeling. Through iteratively updating the distributions of the indicator variables, we down-weight the influence of those suspicious observations. Our model also belongs

[*]*Corresponding author* : J. Sunil Rao

E-mail address : rao.jsunil@gmail.com

to the common contamination model and can be written as

$$(1 - \alpha) * f(x; \theta) + \alpha * g(x) \qquad (2)$$

A subtle but important difference between our method and the other models is that we do not formally specify the spurious distribution $g(x)$ and will not make inference of θ based on "bad" observations. Instead, we approximate $g(x)$ using the average value of the predicted densities at the observations, $\{f(x_i;\hat{\theta})\}$, where $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ with all data. The reasons for this approximation will be explained in Section 2. The conditional posterior distributions of θ and the Bernoulli indicator variables are calculated by using a Gibbs sampler. A summary statistic is derived from the indicator variables' (conditional) posterior distribution which provides a relative importance measure of observations for making inference, and could be used for identifying possible outliers.

Using the simulated data and some real data sets, we will show that our estimates have lower mean square error (MSE) than the ordinary least square (OLS) estimates. Also, we will show that masking has less effect on our method for several data sets when the traditional mean-shift or variance-inflation model fails. Our method can also be extended easily to scenarios like generalized linear regression or multivariate estimation. In that sense, our method will be shown to be quite general in nature.

The paper is organized as follows. Section 2 introduces the general mathematical formulation. Section 3 demonstrates our method with several data sets. In Section 4, the proposed method is summarized and discussed.

2. THE BAYESIAN ROBUST ESTIMATION ALGORITHM

Suppose we have observed data $x_1, x_2, ..., x_n$. Most observations come from distribution $f(x; \theta)$, and a few come from another distribution $g(x)$ with unknown density function. $\underline{x}$ will have a distribution function as Formula 1. The goal is to estimate the unknown parameter θ where θ might be a vector. Suppose we know which distribution each observation comes from and use a latent indicator variable vector $\underline{b} = \{b_1, b_2, ..., b_n\}$ to denote it. If the observation x_i comes from $f(x; \theta)$, then the corresponding $b_i = 1$; if x_i belongs to $g(x)$, then $b_i = 0$. The likelihood function can be written as

$$\prod_{i=1}^{n} f(x_i;\theta)^{b_i} g(x_i)^{1-b_i}$$

Usually $\underline{b}$ is unknown and we can assume that b_i follows a Bernoulli distribution with parameter q, where $q = Prob(b_i = 1)$. Compared with Formula 1, we can see that $q = 1 - \alpha$. Following the Bayesian analysis theme, we assume a prior distribution for θ, and a prior distribution for q. A beta distribution with hyper-parameters (κ, τ) is a common prior choice for q. The joint distribution of all the variables will be

$$\begin{aligned} p(x, \theta, b, q) &= p(x|\theta, b)\, p(b)\, p(\theta)\, p(q) \\ &= \prod_{i=1}^{n} f(x_i;\theta)^{b_i} g(x_i)^{1-b_i} \prod_{i=1}^{n} q^{b_i}(1-q)^{1-b_i} \\ &\quad \times p(\theta) \times q^{\kappa-1}(1-q)^{\tau-1} \end{aligned} \qquad (3)$$

Using $x^{b=1}$ to denote the observations with $b_i = 1$, the conditional posterior distribution of θ is

$$p(\theta | x, b) = p(\theta | x^{b=1}) \propto p(x^{b=1} | \theta)\, p(\theta)$$

The conditional posterior distribution of q is

$$p(q | x, \theta, b) \propto q^{\sum b_i + \kappa - 1}(1-q)^{n - \sum b_i + \tau - 1}$$

which is still a Beta distribution. The conditional posterior distribution of b_i is

$$p(b_i | x, \theta, q) \propto f(x_i;\theta)^{b_i} g(x_i)^{1-b_i} q^{b_i}(1-q)^{1-b_i}$$

which is a Bernoulli distribution with parameter $\dfrac{f(x_i;\theta)q}{f(x_i;\theta)q + (1-q)g(x_i)}$. Since $g(x)$ is unknown and we don't want to assume any functional structure, we need to approximate it non-parametrically. Note that the range of $g(x)$ should be close to that of $f(x; \theta)$. We have tried several ways of approximation including a) a kernel density of x, b) a discrete function $\frac{1}{n}\sum_{i=1}^{n} I_{(x=x_i)} f(x_i;\hat{\theta})$, and c) a delta density $Prob(x = median\ \{f(x_i;\hat{\theta})\}) = 1$. All three methods perform similarly for the simulated and real data sets. For the kernel density, we need to justify the options for the

kernel functional and the smoothing bandwidth. Option b) is a special form of a) and Option c) is a degenerate version of b). We use Option c) from the computing speed consideration. The marginal posterior distribution $b_i|x$ does not have simple analytical form, and we will use the Gibbs sampler to make inference. Below is the general robust estimation algorithm.

Algorithm 1 Bayesian Robust Estimation

Fit the classical model $x \sim f(x; \theta)$, and calculate the predicted densities $\{f(x_i;\hat{\theta})\}$.

$g(x)$ = median $\{f(x_i;\hat{\theta})\}$.

Initially, all $b_i = 1$.

Repeat

Draw θ from $p(\theta \mid x, b)$.

Draw q from $p(q \mid x, \theta, b)$.

Draw b from $p(b \mid x, \theta, q)$.

Monitor the convergence

until The entire distribution is converged

We make inference about θ based on the posterior samples of θ. Since the estimates are based mostly on "good" observations, the inference should be more precise and robust than that of the OLS approach. In the next section, we will show that in linear regression, our estimates have lower mean square error (MSE) than that of OLS in some data analyses. A summary statistic based on the conditional posterior distribution of b_i indicates how often the i^{th} observation is chosen by the model. The observations with less frequency are possible outliers whose roles in estimation are down-weighted.

The function used to estimate the parameter of the conditional posterior distribution of b, $\frac{f*q}{f*q+(1-q)*g}$, is an increasing function of f. Observations that have larger likelihood f, i.e., fitting the model better, are more likely to be selected by the model. If we do not have prior knowledge of the outlier proportion, it is reasonable to assume that the probability of an observation to be outlier is apriori around 0.05 and with high probability to be less than 0.5 (Verdinelli and Wasserman 1991). We suggest the hyper-parameters κ = 3.4, τ = 0.1789, and in a later example, we will show that these values are very robust to situations where there exist large percentages of outliers. The convergence of the Gibbs sampler is monitored by using the multiple sequence method (Gelman and Rubin, 1992). We used 5 parallel Markov chains for our analysis.

3 APPLICATIONS OF THE BAYESIAN ROBUST ESTIMATION APPROACH

In this section, we apply our method to three different problem types: linear regression, logistic regression, and multivariate estimation. The analysis of the simulated data sets and some real data sets show that our approach performs better than other methods.

3.1 Linear Regression

For the linear regression model $Y = X\beta + \varepsilon$ where β is a p-vector regression parameter, the common contamination models for outliers assume a normal mixture distributions for the errors: either the normal variance-inflation model (Box and Tiao 1968; Hoeting *et al*. 1996), $\varepsilon \sim (1-\alpha) * N(0, \sigma^2) + \alpha * N(0, \kappa^2\sigma^2)$ or the normal mean-shift model (Guttman 1973; Abrahan and Box 1978), $\varepsilon \sim (1-\alpha) * N(0, \sigma^2) + \alpha * N(\lambda, \sigma^2)$. We use two simulated data sets and two real data sets to compare the performance of our approach with these two models. Least Median of Squares Regression (LMS) proposed by Rousseeuw (1984) is a widely used robust regression approach. We also fit the data using Splus function "lmsreg", and list the results for the comparison.

Simulation 1: 8 observations from the model $Y = 3 + 2X + N(0, 1)$ and 2 observations from the model $Y = 3+2X + N(0, 3)$ are combined together, and listed in Table 1. These observations are selected so that it can be seen clearly from Fig. 1 that observation 3 and 8 are distinct from others. However, OLS could not "see" the scene and gives biased estimates.

We use the common non-informative prior distribution $f(\beta, \sigma^2) \sim \sigma^2$. For the mean-shift model, we use the prior distribution $N(0, 1)$ for λ as described by Verdinelli and Wasserman (1991). The constant k in the variance inflation model is set as 7 which is suggested by Hoeting *et al*. (1996). The regression estimates and standard errors are summarized in Table 2. Similar to the LMS estimates, our estimates are very close to the true parameter values and the standard

Table 1. The simulated data

X	1.077	2.032	3.018	4.074	5.053	6.054	7.092	8.044	9.038	10.053
Y	5.066	7.812	16.237	9.883	13.595	16.630	17.275	12.414	20.216	23.356

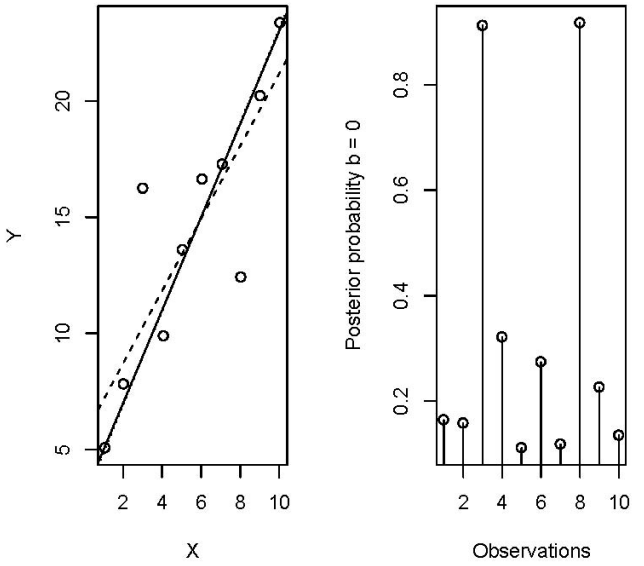

Fig. 1. The simulated data. a) The scatter plot. The solid line has the true intercept and slope. The dashed line is the OLS estimate, and the dotted line is the BRE estimate. b) The conditional posterior probability plot of the indicator variables.

Table 2. Analysis results for the simulated data

Method	Estimate
BRE	$\hat{\beta}$ = (2.837, 2.029) $sd(\hat{\beta})$ = (1.212, 0.807)
OLS	$\hat{\beta}$ = (5.596, 1.558) $sd(\hat{\beta})$ = (2.252, 0.360)
OLS w/o outlier	$\hat{\beta}$ = (3.237, 1.977) $sd(\hat{\beta})$ = (0.714, 0.114)
Mean-Shift	$\hat{\beta}$ = (5.629, 1.555) $sd(\hat{\beta})$ = (2.294, 0.367)
Variance-Inflation	$\hat{\beta}$ = (3.484, 1.944) $sd(\hat{\beta})$ = (0.939, 0.145)
LMS	$\hat{\beta}$ = (3.288, 1.938)

errors are also smaller. The estimates from the ordinary least square (OLS) and mean-shift model are clearly biased. The variance-inflation model works well because the outliers do follow an inflated-variance distribution. From the plot of $p(b_i = 0|x, \theta, q)$ in Fig. 1 b), we can clearly identify that observation 3 and 8 have little weight in estimation.

Stack Loss Data: The first real data set we analyze is the stack loss data (Brownlee 1965), which consists of measurements from a plant for 21 days. The response is the percent of unconverted ammonia that escapes from the plant, which is called stack loss. The predictors include air flow temperature, and acid concentration. This data set has been studied by many statisticians including Daniel and Wood (1980), Atkinson (1985), and Hoeting *et al*. (1996). The common conclusion is that observations 1, 3, 4, and 21 are outliers. In Fig. 2, we plot $p(b_i = 0|x, \theta, q)$ for each observation. These 4 suspicious observations with less weights in modeling are certainly visible. The regression parameter estimates are summarized in Table 3. Our estimates are very close to the OLS estimates without the 4 outliers.

Stars Data: The stars data consists of the effective surface temperature and the light intensity of 47 stars from the star cluster CYG-OB1. The research question of interest is to examine whether there is an appropriate linear relationship between log intensity and log temperature. The data are available in Rousseeuw and Leroy (1987). There are clearly four observations

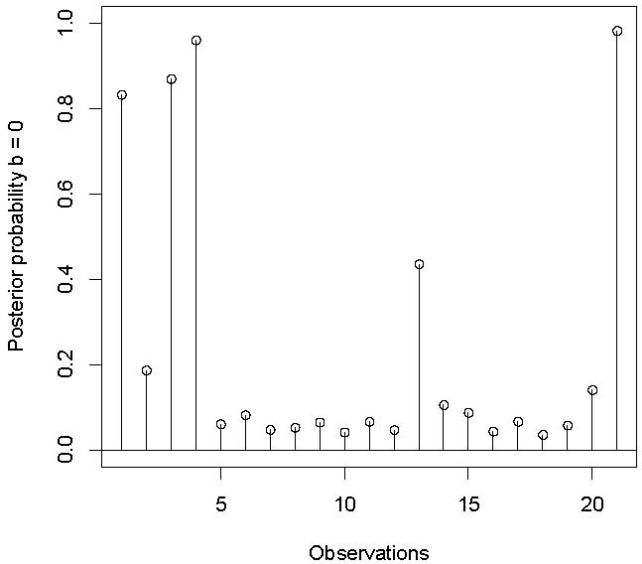

Fig. 2. The conditional posterior probability plot of the indicator variables for the stack loss data

Table 3. Analysis results for the stack loss data

Method	Estimate
BRE	$\hat{\beta}$ = (–37.091, 0.825, 0.536, –0.081) $sd(\hat{\beta})$ = (6.619, 0.118, 0.244, 0.086)
OLS	$\hat{\beta}$ = (–39.920, 0.716, 1.295, –0.152) $sd(\hat{\beta})$ = (11.896, 0.135, 0.368, 0.156)
OLS w/o outlier	$\hat{\beta}$ = (–37.653, 0.798, 0.577, –0.067) $sd(\hat{\beta})$ = (4.732, 0.067, 0.166, 0.062)
Mean-Shift	$\hat{\beta}$ = (–39.108, 0.718, 1.298, –0.165) $sd(\hat{\beta})$ = (11.847, 0.132, 0.373, 0.156)
Variance-Inflation	$\hat{\beta}$ = (–36.809, 0.815, 0.522, –0.074) $sd(\hat{\beta})$ = (4.876, 0.082, 0.210, 0.069)
LMS	$\hat{\beta}$ = (–39.25, 0.75, 0.5, $-6.607 * 10^{-17}$)

(11, 20, 30, 34) separated from others in the scatter plot of Fig. 3, which correspond with giant stars. Justel and Peña (1996) could not identify any outliers using the variance-inflation model and they ascribed the failure to the masking between outliers. Whereas, our conditional posterior plot of *b* in Fig. 3 clearly indicates that these 4 observations are suspicious. As indicated in Table 4, LMS and our approach can detect the correct

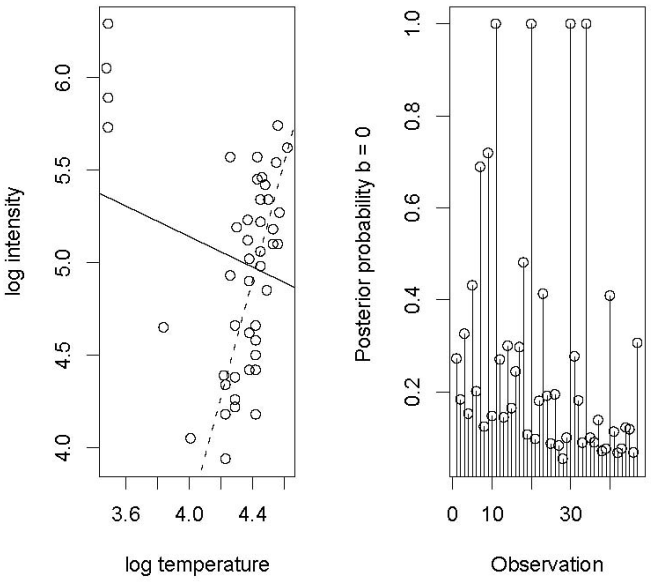

Fig. 3. The stars data. a) The scatter plot. The solid line is the OLS estimate and the dashed line is the BRE estimate. b) The conditional posterior probability plot of the indicator variables.

Table 4. Analysis results for the stars data

Method	Estimate
BRE	$\hat{\beta}$ = (–9.086, 3.181) $sd(\hat{\beta})$ = (3.088, 0.695)
OLS	$\hat{\beta}$ = (6.794, –0.413) $sd(\hat{\beta})$ = (1.237, 0.286)
OLS w/o outlier	$\hat{\beta}$ = (–4.057, 2.047) $sd(\hat{\beta})$ = (1.844, 0.420)
Mean-Shift	$\hat{\beta}$ = (6.721, –0.397) $sd(\hat{\beta})$ = (1.264, 0.293)
Variance Inflation	$\hat{\beta}$ = (6.801, –0.414) $sd(\hat{\beta})$ = (1.201, 0.278)
LMS	$\hat{\beta}$ = (–12.628, 3.971)

direction of the fitted line using the estimate from OLS without outliers as the reference, and other methods fail to do that.

Rousseeuw Data: Another simulated data set we use was generated by Rousseeuw (1984). It consists of fifty 2-dimensional observations. Thirty of them are from model $Y = 2 + X + \varepsilon$, where X's are from a uniform distribution and $\varepsilon \sim N(0, 0.04)$. Another twenty are from a bivariate normal distribution $N((7, 2), 0.5I)$. Because 40% of the data are generated from a different process, which can be seen clearly in Fig. 4, there are heavy masking and swamping effects. It has been used by many researchers as a hard example for robust estimation and outlier detection methods including Justel and Peña (1996, 2001). The variance-inflation model could not identify the outliers due to the masking effect (Justel and Peña, 1996).

This data set is a good example to check the robustness of our hyper-parameter values ($\kappa = 3.4$, $\tau = 0.1789$). We randomly select m observations from the 20 outliers, combine them with the 30 good observations, and apply our method to this new constructed data set. For each m, the whole process is repeated 20 times. When $m \leq 17$, less than 36% of the data are from a different process, our method down-weights outliers correctly each time, and our estimates are very close to the true parameters in the design (Table 5). When $m = 18, 19$, our method converges to

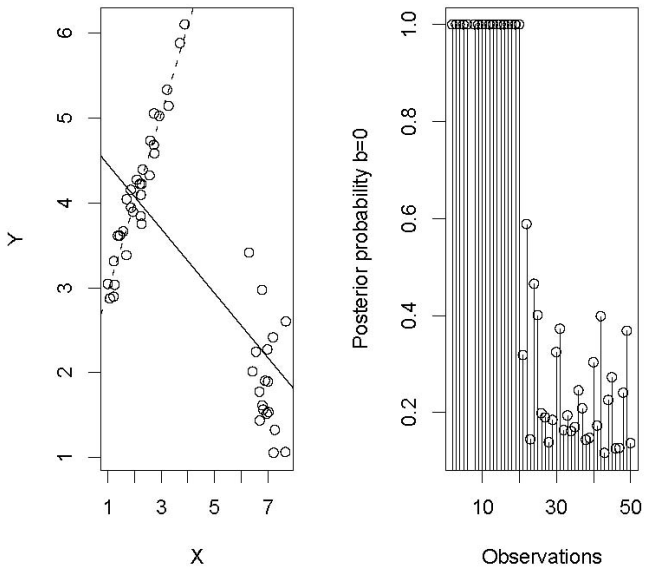

Fig. 4. The Rousseeuw data with 18 outliers. a) The scatter plot. The solid line is the OLS estimate with the full data, and the dashed line is the BRE estimate. b) The conditional posterior probability plot of the indicator variables

Table 5. Analysis results for the Rousseeuw data with 18 outliers

Method	Estimate
BRE	$\hat{\beta}$ = (1.909, 1.044) $sd(\hat{\beta})$ = (0.157, 0.066)
OLS	$\hat{\beta}$ = (4.835, "0.380) $sd(\hat{\beta})$ = (0.274, 0.059)
OLS w/o outlier	$\hat{\beta}$ = (1.880, 1.046) $sd(\hat{\beta})$ = (0.125, 0.054)
LMS	$\hat{\beta}$ = (1.980, 1.040)

the true model 18 out of 20 times. When m = 20, we succeed 10 times. Even though our assumption for the hyper-parameters is that only a small proportion of data are outliers, the values we suggested work well for a wide range of the outlier percentages in this example.

In Table 5, we list the results of our method, OLS and LMS. Ours is very close to that of LMS. The results of the mean-shift model and variance-inflation model are missed since they fail to identify the outliers and get the fitted lines on the wrong direction.

3.1.1 Some Theoretical Results: MSE Comparison

Below we will prove that our estimate of β has less mean square error (MSE) than that of the ordinary least square approach under certain conditions.

Theorem 1 shows that when the random errors follow a normal variance-inflation distribution, the MSE of our estimate is less than that of OLS, which is clear from the analysis results in Section 3.1.

Theorem 1. Suppose $\varepsilon \sim (1-\alpha) * N(0, \sigma^2) + \alpha * N(0, k^2\sigma^2)$, where α is a fixed value. In each MCMC step of our algorithm, we use part of the data to estimate the interested parameters, and suppose the proportion of outliers in this subset is α'. Because that α' could be different for each MCMC step, we assume that α' is a random variable. Let $\hat{\beta}_{bre}$ denote our estimate, $\hat{\beta}_{\alpha'=\xi}$ denote the estimate with α' fixed at a value ξ and let $\hat{\beta}$ denote the OLS estimate, then there exists a value ξ such that MSE($\hat{\beta}_{bre}$) = MSE($\hat{\beta}_{\alpha'=\xi}$). Let tr denote the trace of $(\mathbf{X}^T\mathbf{X})^{-1}$ and tr_b denote the trace of $(\mathbf{X}^{(b=1)^T}\mathbf{X}^{(b=1)})^{-1}$ when α' is fixed at ξ, then MSE($\hat{\beta}_{bre}$) < MSE($\hat{\beta}$) if $\frac{\alpha-\xi}{\alpha} > \frac{tr_b - tr}{tr_b}$.

Theorem 2 shows that when the random errors follow a normal mean-shift distribution, the MSE of our estimate is less than that of OLS, which also can be seen from the analysis results in Section 3.1.

Theorem 2. Suppose $\varepsilon \sim (1-\alpha) * N(0, \sigma^2) + \alpha * N(\lambda, \sigma^2)$, where α is a fixed value. In each MCMC step of our algorithm, we use part of the data to estimate the interested parameters, and suppose the proportion of outliers in this subset is α'. Because that α' could be different for each MCMC step, we assume that α' is a random variable. Let $\hat{\beta}_{bre}$ denote our estimate, $\hat{\beta}_{\alpha'=\xi}$ denote the estimate with α' fixed at a value ξ and let $\hat{\beta}$ denote the OLS estimate, then there exists a value ξ such that MSE($\hat{\beta}_{bre}$)= MSE($\hat{\beta}_{\alpha'=\xi}$). Let tr denote the trace of $(\mathbf{X}^T\mathbf{X})^{-1}$ and tr_b denote the trace of $(\mathbf{X}^{(b=1)^T}\mathbf{X}^{(b=1)})^{-1}$ when α' is fixed at ξ, then MSE($\hat{\beta}_{bre}$) < MSE($\hat{\beta}$) if

$$\frac{\alpha(1-\alpha) - \xi(1-\xi)}{\alpha(1-\alpha)} > \frac{tr_b - tr}{tr_b}$$

Theorem 3 shows that when the random errors follow a normal mean-shift distribution, the estimate of the normal mean-shift model has larger MSE than that of OLS, even it correctly specifies the distribution of the outliers.

Theorem 3. Suppose $\varepsilon \sim (1-\alpha) * N(0, \sigma^2) + \alpha * N(\lambda, \sigma^2)$, and $\hat{\beta}_{M-S}$ is the estimate using the normal mean-shift model, then the MSE($\hat{\beta}_{M-S}$) > MSE($\hat{\beta}^{OLS}$).

3.2 Generalized Linear Regression

Another example data we use are from Brown (1980), which consists of the measurements of 53 prostatic cancer patients. Each patient has six measurements taken: age, acid level, X-ray result, tumor size, tumor grade, and nodal involvement. The research question is to explore the relationship between the binary response of nodal involvement and the other variables. This data set has been analyzed by Collett (1991) and Albert and Chib (1995). Albert and Chib considered the model with four covariates log(acid), X-ray, size, and grade. With a Bayesian residual analysis, they claimed that observations 9, 26, 35, 37 are outliers.

Using a logistic regression model, the likelihood function is

$$p(y \mid x, \beta, \theta) = \prod_{i=1}^{n} [(\frac{e^{x_i\beta}}{1+e^{x_i\beta}})^{y_i} (\frac{1}{1+e^{x_i\beta}})^{1-y_i}]^{b_i} g(x_i)^{1-b_i} \tag{4}$$

We assume a non-informative prior $P(\beta) = const$, and fit Algorithm 1 use the same beta prior for b_i as other examples. As shown in Fig. 5, observations 37, 9, 26, 35 appear to be different from the others. The estimates are summarized in Table 6. Again, our parameter estimates of interest are closer to those calculated without suspicious observations.

3.3 Mulitvariate Estimation

This example is a two dimensional density estimation problem described in Barnett and Lewis (1994, p. 289). It consists of the yields of grass on two totally untreated plots at Rothamsted Experimental Station for the 50 years from 1941 to 1990. The purpose is to estimate the average yield and tests the discordance

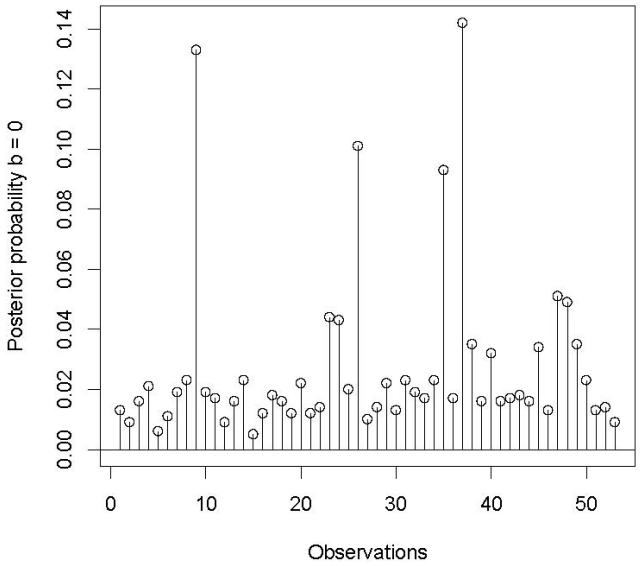

Fig. 5. The conditional posterior probability plot of the indicator variables for the prostatic data

Table 6. Analysis results for the prostatic data

Method	Estimate
BRE	$\hat{\beta}$ = (–1.382, 2.694, 2.182, 1.662, 0.844) $sd(\hat{\beta})$ = (0.856, 1.321, 0.992, 0.928, 0.854)
GLM	$\hat{\beta}$ = (–1.306, 2.512, 2.011, 1.544, 0.851) $sd(\hat{\beta})$ = (0.727, 1.173, 0.821, 0.780, 0.775)
GLM w/o outlier	$\hat{\beta}$ = (–3.263, 4.730, 4.593, 3.604, 1.827) $sd(\hat{\beta})$ = (1.365, 1.996, 1.784, 1.570, 1.151)

of several suspicious observations. Following the assumption of Barnett and Lewis (1994) and Varbanov (1998), we assume that observations x have a multivariate normal distribution $X|\mu, \Sigma \sim N(\mu, \Sigma)$. The commonly proposed non-informative prior distribution $p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$ and the same beta prior as other examples are used. Using Algorithm 1, the first two observations with the lowest frequencies being selected by the model are case 25 for year 1965 and 31 for year 1971. The result coincides with the results of Barnett and Lewis (1994) and Varbanov (1998). The estimates are summarized in Table 7. Compared with the estimates using all data, our estimates are much closer

Table 7. Analysis results for the barnett and lewis data

Method	BRE	MLE	MLE w/o outlier
Estimates	$\hat{\mu}$ = (1.1340, 1.3347) $\hat{\Sigma} = \begin{pmatrix} 0.1350 & 0.1253 \\ 0.1253 & 0.1648 \end{pmatrix}$	$\hat{\mu}$ = (1.1740, 1.4492) $\hat{\Sigma} = \begin{pmatrix} 0.1877 & 0.1996 \\ 0.1996 & 0.3392 \end{pmatrix}$	$\hat{\mu}$ = (1.1344, 1.3765) $\hat{\Sigma} = \begin{pmatrix} 0.1390 & 0.1277 \\ 0.1277 & 0.2158 \end{pmatrix}$

to those calculated without suspicious observations. Following Theorem 4 shows that our estimate has less MSE than the estimate with all data when the data follow a contaminated distribution.

Theorem 4. Suppose $X \sim (1 - \alpha) * N(\mu_1, \sigma_1^2) + \alpha * N(\mu_2, \sigma_2^2)$, where α is a fixed value. In each MCMC step of our algorithm, we use part of the data to estimate the interested parameters, and suppose the proportion of outliers in this subset is α'. Because that α' could be different for each MCMC step, we assume that $\alpha'$is a random variable. Let $\hat{\mu}_{bre}$ denote our estimate for the location parameter, $\hat{\mu}_{\alpha'=\xi}$ denote the estimate with α' fixed at ξ, and let $\hat{\mu}$ denote the estimate with all data, then there exists a value ξ such that MSE($\hat{\mu}_{bre}$) = MSE($\hat{\mu}_{\alpha'=\xi}$), MSE($\hat{\mu}_{bre}$) < MSE($\hat{\mu}$) if $\xi < \alpha$.

4. DISCUSSIONS

We invent an indicator variable for each observation and devise a method to down-weight "bad" observations in the estimation. The conditional posterior distributions of the indicator variables are used to locate possible outliers. It can be easily applied to many statistical models with an explicit likelihood function. While our method is similar to the variance inflation model and mean-shift model, there are important differences. We do not assume a special distribution for the error term and we do not use the possible outliers for modeling.

The underlying assumption is that only few observations are outliers, and we choose the hyper-parameters based on this assumption. With the Rousseeuw data analysis, we show that the suggested hyper-parameters are quite robust. The conditional posterior plot of the indicator variables could be used as an index for the relative importance measure of the observations and potential outlier identification. However, we emphasize that there is no critical criteria for outliers and the investigation of suspicious observations should be conducted carefully.

Our method can also be applied to other generalized linear models such as the partial likelihood function in survivor data analysis. We are also exploring ways of combining our approach with some variable selection methods to do variable selection and outlier detection simultaneously. Some candidate methods are SSVS (George and McCulloch 1993) and the spike and slab model (Ishwaran and Rao 2005).

ACKNOWLEDGEMENTS

Guan Xing is a Sr. Reserach Biostatistician at Bristol-Myers Squibb, and this work comprised a portion of his dissertation. J. Sunil Rao is Professor of Biostatistics and Genetic Epidemiology at Case Western Reserve University. His research is partially supported by NSF grant DMS-0203724. The authors would like to thank Joe Sedransk for helpful conversations while conducting this research.

REFERENCES

Abraham, B. and Box, G.E.P. (1978). Linear models and spurious observations. *Appl. Statist.*, **27**, 131-138.

Albert, J. and Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, **4**, 747-759.

Atkinson, A.C. (1985). *Plots, Transformations and Regression*. Clarendon Press, Oxford.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. 3rd ed. John Wiley & Sons, Chichester.

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

Brown, B.W. (1980). Prediction analysis for binary data. In: *Biostatistics Casebook*, (R.J. Miller, B. Efron, B.W. Brown and L.E. Moses eds.), pp. 3-18. Wiley, New York.

Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*. 2nd ed., Wiley, New York.

Collett, D. (1991). *Modeling Binary Data*. Chapman and Hall, London.

Daniel, C. and Wood, F.S. (1980). *Fitting Equations to Data.* Wiley, New York.

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

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*. 2nd ed., Chapman and Hall.

George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, **88**, 881-889.

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

Hoeting, J., Raftery, A.E. and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Comput. Statist. Data Anal.*, **22**, 251-270.

Ishwaran, H. and Rao, J.S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.*, **33**, 730-773.

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

Justel, A. and Peña, D. (2001). Bayesian unmasking in linear models. *Comput. Statist. Data Anal.*, **36**, 69-84.

Rousseeuw, P.J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871-880.

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

Varbanov, A. (1998). Bayesian approach to outlier detection in multivariate normal samples and linear models. *Comm. Statist.-Theory Methods*, **27**, 547-557.

Verdinelli, I. and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler. *Statist. Comput.*, **1**, 105-117.

APPENDIX

Proof of Theorem 1

We will use the first mean value theorem for integration:

If $f:[a, b] \to R$ is a continuous function and $g:[a, b] \to R$ an integrable positive function, then there exists a number ξ in $[a, b]$ such that

$$\int_a^b f(x)g(x)\mathrm{d}x = f(\xi)\int_a^b g(x)\mathrm{d}x$$

For the normal variance-inflation model, the mean square error for the ordinary least square (OLS) estimate of β is $E(\hat{\beta}-\beta)^T(\hat{\beta}-\beta) = [(1-\alpha)\sigma^2 + \alpha k^2\sigma^2]\, tr((X^TX)^{-1})$. If $f(\alpha')$ is the MSE function when there are α' proportion of outliers, $g(\alpha')$ is the density function of α', $[a, b]$ is the domain of α', then the MSE of our estimate would be $\int_a^b f(\alpha')g(\alpha')\mathrm{d}\alpha'$.

Based on the mean value theorem for integration, there exists $\xi \in [a, b]$ such that MSE($\hat{\beta}_{bre}$) equals to $[(1-\xi)\sigma^2 + \xi k^2\sigma^2]tr((X^{b=1^T}X^{b=1})^{-1})$.

$$\begin{aligned}\mathrm{MSE}(\hat{\beta}) - \mathrm{MSE}(\hat{\beta}_{bre}) &= (\alpha-\xi)tr_b(k^2-1)\sigma^2 \\ &\quad + \sigma^2(1+\alpha(k^2-1))(tr-tr_b) \\ &= \left[\frac{\alpha-\xi}{\alpha} - \frac{1+\alpha(k^2-1)}{\alpha(k^2-1)} * \frac{tr_b-tr}{tr_b}\right] * \alpha(k^2-1)\sigma^2 tr_b\end{aligned}$$

So that MSE($\hat{\beta}$) > MSE($\hat{\beta}_{bre}$) when

$$\frac{\alpha-\xi}{\alpha} > \frac{1+\alpha(k^2-1)}{\alpha(k^2-1)} * \frac{tr_b-tr}{tr_b}$$

The first term on the right of the equation is close to 1 when k is large. Hence, MSE($\hat{\beta}$) > MSE($\hat{\beta}_{bre}$) when $\frac{\alpha-\xi}{\alpha} > \frac{tr_b-tr}{tr_b}$.

$\xi < \alpha$ because we choose observations based on the likelihood and observations with large errors have less chance to be selected. We could make the difference between $tr((X^TX)^{-1})$ and $tr((X^{(b=1)^T}X^{(b=1)})^{-1})$ small by using a large prior probability parameter of b.

Proof of Theorem 2

For the normal mean-shift model, the mean square error for the ordinary least square estimate of $\hat{\beta}$ is $E(\hat{\beta}-\beta)^T(\hat{\beta}-\beta) = [\sigma^2 + \lambda^2\alpha(1-\alpha)]\, tr((X^TX)^{-1})$. As we did in the proof of Theorem 1, we can prove that there exists ξ such that the mean square error for our estimate $\hat{\beta}_{bre}$ is $[\sigma^2 + \lambda^2\xi(1-\xi)]tr((X^{b=1^T}X^{b=1})^{-1})$.

$$\begin{aligned}MSE(\hat{\beta}) - MSE(\hat{\beta}_{bre}) &= [\alpha(1-\alpha) - \xi(1-\xi)]\, tr_b\lambda^2 \\ &\quad - [\sigma^2 + \lambda^2\alpha(1-\alpha)](tr_b - tr) \\ &= \left[\frac{\alpha(1-\alpha)-\xi(1-\xi)}{\alpha(1-\alpha)} - \frac{\sigma^2+\lambda^2\alpha(1-\alpha)}{\lambda^2\alpha(1-\alpha)} * \frac{tr_b-tr}{tr_b}\right] * \alpha(1-\alpha)\lambda^2 tr_b\end{aligned}$$

So that MSE($\hat{\beta}$) > MSE($\hat{\beta}_{bre}$) when

$$\frac{\alpha(1-\alpha)-\xi(1-\xi)}{\alpha(1-\alpha)} > \frac{\sigma^2+\lambda^2\alpha(1-\alpha)}{\lambda^2\alpha(1-\alpha)} * \frac{tr_b-tr}{tr_b}$$

The first term on the right of the equation is close to 1 when λ is large. Hence, MSE($\hat{\beta}$) > MSE($\hat{\beta}_{bre}$) when $\frac{\alpha(1-\alpha)-\xi(1-\xi)}{\alpha(1-\alpha)} > \frac{tr_b-tr}{tr_b}$.

Proof of Theorem 3

For the normal mean-shift model, the mean square error for the ordinary least square estimate is $E(\hat{\beta}-\beta)^T(\hat{\beta}-\beta) = [\sigma^2 + \alpha(1-\alpha)\lambda^2]\, tr(((X^TX)^{-1})$ which is $\mathrm{var}(\varepsilon) * tr((X^TX)^{-1})$. Mean-shift model corrects y_i with $y_i^* = y_i - \lambda$, when the i^{th} observation is assumed as an outlier. Suppose that the mean-shift model claims γ proportion of the data are outliers, we have a $1-\gamma$ proportion of errors that still have the mean-shift

mixture distribution and a γ proportion of errors that have the mixture distribution $(1 - \alpha)\ N\ (-\lambda,\ \sigma^2) + \alpha * N(0,\ \sigma^2)$.

Combining them together, the error term has a 3-component mixture distribution.

$$\varepsilon_{M-S} \sim (1 - \alpha - \gamma + 2\alpha\gamma)\ N\ (0,\ \sigma^2) + \alpha(1 - \gamma)\ N\ (\lambda,\ \sigma^2) + (1 - \alpha)\ \gamma\ N(-\lambda,\ \sigma^2)$$

$E(\varepsilon_{M-S}) = \alpha(1 - \gamma)\lambda - (1 - \alpha)\ \gamma\lambda = (\alpha - \gamma)\lambda$, and

$$E(\varepsilon_{M-S}^2) = (1 - \alpha - \gamma + 2\alpha\gamma)\sigma^2 + \alpha\ (1 - \gamma)\ (\lambda^2 + \sigma^2) + (1 - \alpha)\ \gamma(\lambda^2 + \sigma^2)$$

So that we can get

$$Var(\varepsilon_{M-S}) = E(\varepsilon^2) - [E(\varepsilon)]^2 = \sigma^2 + (\alpha(1-\alpha) + \gamma(1-\gamma))\lambda^2,$$

which is larger than $Var(\varepsilon)$. Hence, $\mathrm{MSE}(\hat{\beta}_{M-S}) > \mathrm{MSE}(\hat{\beta}_{OLS})$.

Proof of Theorem 4

$X \sim (1 - \alpha)\ \mathrm{N}(\mu_1,\ \sigma_1^2) + \alpha * \mathrm{N}(\mu_2,\ \sigma_2^2)$, so we have $\bar{X} \sim (1 - \alpha)\ \mathrm{N}(\mu_1,\ \sigma_1^2/n) + \alpha * \mathrm{N}(\mu_2,\ \sigma_2^2/n)$. $\mathrm{MSE}(\hat{\mu}) = \mathrm{MSE}(\bar{X}) = (E\bar{X} - \mu_1)^2 + \mathrm{Var}(\bar{X}) = (1 - \alpha)\sigma_1^2/n + \alpha * \sigma_2^2/n + \alpha(\mu_1 - \mu_2)^2$. Let $\hat{\mu}_{bre}$ denote our estimate, as we did in the proof of Theorem 1 and 2, we can show that there exists ξ such that $\mathrm{MSE}(\hat{\mu}_{bre}) = (1 - \xi)\sigma_1^2/n + \xi * \sigma_2^2/n + \xi(\mu_1 - \mu_2)^2$. $\mathrm{MSE}(\hat{\mu}) - \mathrm{MSE}(\hat{\mu}_{bre}) = (\alpha - \xi)\ [\sigma_2^2/n - \sigma_1^2/n + (\mu_1 - \mu_2)^2]$. $\sigma_2^2 \geq \sigma_1^2$ is the general assumption of the robust models.