ELSEVIER Computational Statistics & Data Analysis 22 (1996) 251–270

COMPUTATIONAL STATISTICS & DATA ANALYSIS

A method for simultaneous variable selection and outlier identification in linear regression *

Jennifer Hoeting [a,*,1], Adrian E. Raftery [b,1], David Madigan [b,2]

[a] *Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA*
[b] *Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195, USA*

Received March 1995; revised September 1995

Abstract

We suggest a method for simultaneous variable selection and outlier identification based on the computation of posterior model probabilities. This avoids the problem that the model you select depends upon the order in which variable selection and outlier identification are carried out. Our method can find multiple outliers and appears to be successful in identifying masked outliers.

We also address the problem of model uncertainty via Bayesian model averaging. For problems where the number of models is large, we suggest a Markov chain Monte Carlo approach to approximate the Bayesian model average over the space of all possible variables and outliers under consideration. Software for implementing this approach is described. In an example, we show that model averaging via simultaneous variable selection and outlier identification improves predictive performance and provides more accurate prediction intervals as compared to any single model that might reasonably be selected.

Keywords: Bayesian model averaging; Markov chain Monte Carlo model composition; Masking; Model uncertainty; Posterior model probability

1. Introduction

Many approaches for the selection of variables and the identification of outliers have been proposed. Most authors focus on these problems separately. Adams (1991) and Blettner and Sauerbrei (1993) pointed out that the model that is selected depends

* Corresponding author. E-mail: jah@lamar. colostate.edu.

[1] Partially supported by ONR Contract N-00014-91-J-1074.
[2] Partially supported by NSF grant no. DMS 92111627.

SSDI 0167-9473(95)00053-4

upon the order in which variable selection and outlier identification are performed. In this paper we define a model as a set of variables and a set of observations identified as outliers.

Another difficulty in outlier identification is masking, where multiple outliers in a data set conceal the presence of additional outliers. Several authors have suggested methods to overcome masking, including Atkinson (1986a) and Hadi (1992), but these methods typically involve removing the 'masking' outliers from the data set before the 'masked' outliers can be identified.

We offer a simultaneous approach to variable selection and outlier identification based on Bayesian posterior model probabilities. "Simultaneous Bayesian variable selection and outlier identification" (SVO) overcomes the problem that order of methods influences the choice of outliers and variables. SVO includes a method for identifying multiple outliers which appears to be successful in the identification of masked outliers.

We also consider the problem of model uncertainty in linear regression. The typical approach to model selection involves choosing a single set of variables and identifying a single set of observations as outliers. Subsequent inferences ignore uncertainty involved in the selection of the model. A complete Bayesian solution to this problem involves averaging over *all* possible models when making inferences about quantities of interest. Indeed, Bayesian model averaging provides optimal predictive ability (Madigan and Raftery, 1994). In many applications however, this approach will not be practical due to the large number of models for which posteriors need to be computed.

To overcome this problem we suggest a Markov chain Monte Carlo approach to approximate the Bayesian model average for the space of all possible variables and outliers under consideration. Markov chain Monte Carlo model composition (MC^3) was originally proposed by Madigan and York (1995) and was adapted for linear regression models by Raftery et al. (1994). We show in an example that model averaging via MC^3 provides better predictive performance than any single model which might reasonably have been selected. Software for implementing MC^3 is described.

In the next section we discuss various approaches to outlier identification. In Section 3 we outline our method for SVO including our method for the identification of multiple outliers. In Section 4 we provide two examples using SVO. In Section 5 we summarize Bayesian model averaging and outline MC^3 as implemented for the simultaneous approach. We also discuss the assessment of predictive performance and provide an example comparing the predictive performance of BMA to the predictive performance of single models that would have been chosen using standard techniques. Conclusions are given in Section 6. In the appendix we describe software for implementing MC^3.

2. Outliers in linear regression

Observations that do not follow the same model as the rest of the data are typically called outliers. There is a vast literature on methods for handling outliers including

at least three books (Rousseeuw and Leroy, 1987; Barnett and Lewis 1994; Hawkins, 1980). Outliers are typically modeled by either a shift in mean (i.e., for an outlier y_i, let $y_j = X_j\beta + \varepsilon_j$ for all $j \neq i$ and let $y_i = X_i\beta + \delta + \varepsilon_i$) or via a shift in variance (described in Section 3.1). The mean-slippage model is typically used to *identify* outliers to make them available for further study. The variance–inflation model is often adopted for robust techniques with the aim of tolerating or *accommodating* outliers. We have adopted the variance–inflation model in this work.

Many methods have been suggested for detecting single outliers. For a comparison of many of the available methods, see Chatterjee and Hadi (1986). Bayesian outlier models have also been much discussed in the literature, including Box and Tiao (1968), Guttman et al. (1978), Verdinelli and Wasserman (1991), and Pettit (1992).

If a data set has multiple outliers, then the outliers may mask one another making outlier identification difficult. If masked outliers are not removed from the model as a group, their presence goes undetected. An obvious solution to this problem, the computationally intensive task of consideration of all subsets of observations to be potential outliers, is typically impossible to carry out due to the large number of subsets to be considered.

Several authors have suggested algorithms for detecting multiple outliers including Hadi (1990), Kianifard and Swallow (1989) and Marasinghe (1985). There are also robust methods which produce coefficient estimates that are consistent with the majority of the data. These include work by Rousseeuw (1984), Heiberger and Becker (1991), and Bloomfield and Steiger (1983).

The method we use to identify multiple outliers involves two steps. In a first exploratory step we use a robust technique to identify a set of potential outliers. The robust approach typically identifies a large number of potential outliers. In the second step, we compute all possible posterior model probabilities or use MC^3, considering all possible subsets of the set of potential outliers. This two–step method is computationally feasible, and it allows for groups of observations to be considered simultaneously as potential outliers. In the examples we have considered to date, our method successfully identifies masked outliers. We describe the method in detail below.

3. Simultaneous variable selection and outlier identification

3.1. Bayesian framework and selection of prior distributions

We adopt a variance–inflation model for outliers as follows: Let $Y = X\beta + \varepsilon$ where the observed data on the predictors are contained in the $n \times (p+1)$ matrix X and the observed data on the dependent variable are contained in the n-vector Y. We assume that the ε's in distinct cases are independent where

$$\varepsilon \sim \begin{cases} \mathrm{N}(0, \sigma^2) & \text{w.p. } (1-\pi) \\ \mathrm{N}(0, K^2\sigma^2) & \text{w.p. } \pi. \end{cases} \tag{1}$$

Here π is the probability of an outlier and K^2 is the variance–inflation parameter.

We typically consider all models equally likely a priori and the $(p+1)$ parameter vector β and σ^2 to be unknown. Where possible, informative prior distributions for β and σ^2 should be elicited and incorporated into the analysis – see Kadane et al. (1980) and Garthwaite and Dickey (1992). In the absence of expert opinion we seek prior distributions which reflect uncertainty about the parameters and also embody reasonable a priori constraints. We use prior distributions that are proper but reasonably flat over the range of parameter values that could plausibly arise. These represent the common situation where there is some prior information, but rather little of it. We use the standard normal-gamma conjugate class of priors,

$$\beta \sim \mathrm{N}(\mu, \sigma^2 V),$$

$$\frac{\nu\lambda}{\sigma^2} \sim \chi^2_\nu.$$

Here ν, λ, the $(p+1)\times(p+1)$ matrix V and the $(p+1)$-vector μ are hyperparameters to be chosen.

For non-categorical predictor variables we assume the individual β's to be independent a priori. We center the distribution of β on zero (apart from β_0) and choose $\mu = (\hat{\beta}_0, 0, 0, \ldots, 0)$ where $\hat{\beta}_0$ is the ordinary least-squares estimate of β_0. The covariance matrix V is diagonal with entries $(s_Y^2, \phi^2 s_1^{-2}, \phi^2 s_2^{-2}, \ldots, \phi^2 s_p^{-2})$ where s_Y^2 denotes the sample variance of Y, s_i^2 denotes the sample variance of X_i for $i = 1, \ldots, p$, and ϕ is a hyperparameter to be chosen. The prior variance of β_0 is chosen conservatively and represents an upper bound on the reasonable variance for this parameter. The variances of the remaining β-parameters are chosen to reflect increasing precision about each β_i as the variance of the corresponding X_i increases and to be invariant to scale changes in both the predictor variables and the response variable. For details of our treatment of categorical predictor variables, see Hoeting (1994).

The marginal distribution of the response y based on the proper priors discussed above is a non-central Student's t-distribution with ν degrees of freedom, mean $X\mu$, and variance $[\nu/(\nu-2)]\lambda(\Sigma + XVX^t)$ where Σ is a diagonal matrix with K^2 on the diagonal for observations identified to be outliers and 1's elsewhere.

3.2. *Choosing hyperparameter values for the prior distributions*

Below we briefly describe the rationale behind our choice of the hyperparameters ν, λ, ϕ, π, and K.

We consider the outlier hyperparameters, π and K, separately from the regression hyperparameters, ν, λ, and ϕ. To choose the regression hyperparameters we define a number of reasonable desiderata and attempt to satisfy them. In what follows we assume that all the variables have been standardized to have zero mean and sample variance one. We would like:

1. The prior density $p(\beta_1, \ldots, \beta_p)$ to be reasonably flat over the unit hypercube $[-1, 1]^p$.

2. $p(\sigma^2)$ to be reasonably flat over $(a, 1]$ for some small a.
3. $\Pr(\sigma^2 \leq 1)$ to be large.

The order of importance of these desiderata is roughly the order in which they are listed.

To choose an appropriate value of a in desideratum 2 may require consideration of the data. The values of R^2 for the most likely models should not exceed $(1-a)$ by much. More generally, we may replace the interval $(a, 1]$ in desideratum 2 by the interval (a, b), where $b < 1$ and the values of R^2 for the plausible models are between $(1-b)$ and $(1-a)$. This is to avoid an undue influence of prior tail behavior on the result.

Before using the variable selection/outlier identification framework described below, we recommend that R^2 be computed for the full model (with no outliers identified). If R^2 is less than 0.9, then we suggest using the hyperparameter values $\nu = 2.58$, $\lambda = 0.28$, and $\phi = 2.85$ (hyperparameter set 1). If R^2 is high ($R^2 \geq 0.9$), we suggest using the hyperparameter values $\nu = 0.2$, $\lambda = 0.1684$, and $\phi = 9.20$ (hyperparameter set 2).

The outlier hyperparameters, π and K, have easily definable roles in the model with π defined as the proportion of outliers and K defined as the variance–inflation parameter. In the examples we assume fixed values for the hyperparameters π and K. An analyst may have a prior notion as to what these values should be before looking at the data. Increasing K will decrease the influence of an outlying observation on the posterior model probability and decreasing K should have the opposite effect. Since a variance–inflation parameter of 7 has been found reasonable in other contexts (e.g. Taplin and Raftery, 1994), we have chosen to use the value $K = 7$ for our analyses.

Increasing π, the prior parameter for the proportion of outliers, corresponds to an increase in the likelihood that an individual observation will be identified as an outlier. For small data sets ($n < 50$), we suggest setting the proportion of outliers, π, equal to 0.1 and for larger data sets we use $\pi = 0.02$. While this choice may appear somewhat arbitrary, this setup allows the user to assume a priori that, on average, there is at least one outlier in a data set with more than 10 observations. We have found some sensitivity of the results to the values of π in that increasing the value of π increases the posterior probability for individual outliers.

3.3. Masking

To overcome masking, we use least median of squares (LMS) regression (Rousseeuw, 1984), to prescreen the data. The aim is to identify all the potential outliers in this initial pass. Atkinson (1986a) uses LMS regression to prescreen the data in a similar manner. We chose to use LMS regression because it has a very high breakdown point (close to 1/2) and tends to identify large numbers of observations as outliers, thus minimizing the chance that an outlier will be missed at this stage. It should be noted that LMS regression can be locally unstable. Hettmansperger and

Sheather (1992) demonstrate that small changes in centrally located data can result in large changes in LMS estimates. The user of the technique described below is advised to examine the results to determine whether the set of potential outliers is reasonable.

We use the following procedure to identify potential outliers:

1. Perform LMS regression on the full data set. We use the "lmsreg" function in S-PLUS.®

2. Compute a robust scale estimate of the residuals from step 1. We use 1.4826 times the median absolute deviation, which is a consistent estimator of the standard deviation for Gaussian data (Hoaglin et al., 1983).

3. Compute standardized residuals by dividing the residuals by the robust scale estimate from step 2.

4. All observations such that the absolute value of the standardized residual is greater than some threshold δ are considered to be potential outliers. In the examples below we use $\delta = 2$.

This prescreening procedure produces a conservative (i.e., large) list of potential outliers. We consider all possible combinations of this conservative list as potential outliers for SVO. In the examples we have examined to date, this method overcomes masking while avoiding the consideration of an impossibly large number of groups of potential outliers.

4. Examples

In the two examples below, we demonstrate our simultaneous approach to variable selection and outlier identification.

4.1. Scottish Hill Racing

The first example involves data supplied by the Scottish Hill Runners Association (Atkinson 1986b)[3]. The purpose of the study is to investigate the relationship between record time of 35 hill races and two predictors: distance is the total length of the race, measured in miles, and climb is the total elevation gained in the race, measured in feet. One would expect that longer races and larger climbs would be associated with longer record times (Fig. 1).

Several authors have examined these data using both predictors in their analyses. Atkinson (1986b) and Hadi (1992) concluded that races 7 and 18 are outliers. After they removed observations 7 and 18, their methods indicated that observation 33 is also an outlier. Thus, observations 7 and 18 mask observation 33. After race numbers 7, 18, and 33 are removed from the data, standard diagnostic checking (e.g., Weisberg, 1985) does not reveal any gross violations of the assumptions underlying normal linear regression.

[3] All data used in this paper are available on the World Wide Web at the URL `http://www.stat.colostate.edu/~jah/index.html`.

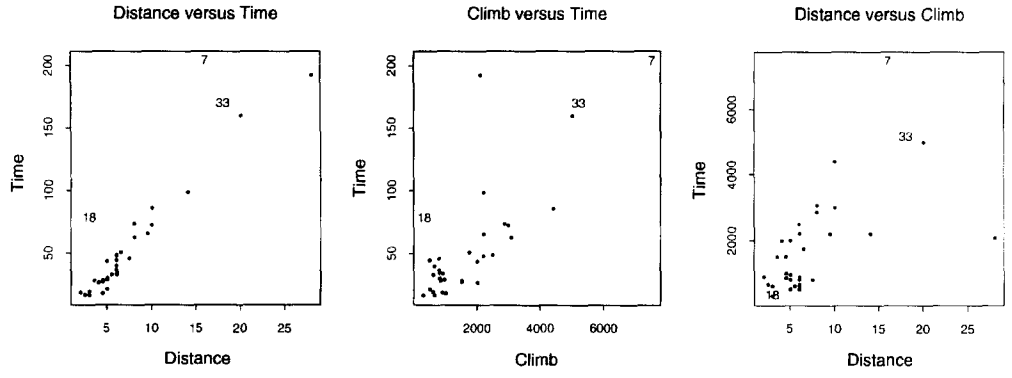

Fig. 1. Scatter plots of Scottish Hill Racing data. Numbers correspond to race numbers 7, 18, 33. Distance is givin in miles, time is given in minutes, climb is given in feet.

Table 1
Races data: models with the 10 highest posterior model probabilities. All models included the variables Distance and Climb

Outliers (race number)	Posterior model prob. (%)
7 18 33	56
7 18 19 33	11
7 14 18 33	6
6 7 18 33	5
7 18	4
7 11 18 33	2
7 18 26 33	2
7 15 18 33	1
7 10 18 33	1
6 7 18 19 33	1

We used the method described in Section 3.3 to identify potential outliers. The prescreening procedure indicated 12 races (races 6, 7, 10, 11, 14, 15, 17, 18, 19, 26, 33, and 35) as potential outliers.

Since $R^2 > 0.9$ and $n < 50$, we used hyperparameter set 2 and $\pi = 0.1$, $K = 7$. Using the set of 12 potential outliers identified by the prescreening procedure, we calculated the posterior model probabilities for the $2^2 \times 2^{12}$ combinations of variables and outliers.

The posterior probability that the coefficients for the predictors climb and distance are non-zero is close to 100%. The models with the 10 highest posterior model probabilities are shown in Table 1. The model with races 7, 18, and 33 as outliers has a posterior model probability of 56%.

The outlier posterior probability for each observation identified to be a potential outlier is given in Table 2. The outlier posterior probability for observation i is the sum of the posterior probabilities of models in which observation i is an outlier. Race

Table 2
Races data: outlier posterior probability for each potential outlier, expressed as a percentage. The outlier posterior probability for observation i is the sum of the posterior probability across models in which observation i is identified as an outlier

Race no.	6	7	10	11	14	15	17	18	19	26	33	35
Outlier p.p.	8	100	2	3	9	2	2	100	16	3	94	2

number 33 was labeled an outlier in models accounting for 94% of the outlier posterior model probability. Races 7 and 18 were labeled outliers in models accounting for nearly 100% of the outlier posterior probability. These results provide strong evidence that races 7, 18, and 33 are aberrant in some way and should be further investigated.

The results are somewhat sensitive to changes in the values of the regression hyperparameters (ν, λ, and ϕ) due to the tail behavior of the prior for σ^2. When hyperparameter set 1 is used, our Bayesian framework does not provide strong evidence that observation 33, which is masked by observations 7 and 18, is an outlier. Section 3.2 provides guidelines for choosing hyperparameters in this and similar situations. The results are not sensitive to reasonable changes in the values of the outlier hyperparameters (K and π).

The Bayesian framework we use here to identify outliers performs better than the methods of Hadi (1992) and Atkinson (1986a) in the sense that it identifies all three observations (7, 18, 33) as outlying at one time. Atkinson (1986a) used a similar prescreening procedure to identify potential outliers, but his method identifies observations 7 and 18 as outliers in a second pass over the data before a third pass where he identifies the masked outlier, observation 33.

In conclusion, both race and climb are important predictors of record time for Scottish Hill races. There is strong evidence that races 7, 18, and 33 are outlying. With a total climb of 7500 ft, race number 7 has the largest total elevation gain of any race. Similarly, race number 33 has the second longest climb of any race and is the third longest race. In a more recent analysis of these data, Atkinson (1988) reports that the time for race number 18 is incorrect. An anonymous referee noted that the correct time for race 18 should be 16 min, 7 s as reported to him by Geoff Cohen of the University of Edinburgh. The original data were used here so that results could be compared with the results of Hadi and Atkinson.

4.2. *Stack loss*

The stack loss data (Brownlee, 1965) consist of 21 days of operation from a plant for the oxidation of ammonia as a stage in the production of nitric acid. The response is called "stack loss" which is the percent of unconverted ammonia that escapes from the plant. There are three explanatory variables (Fig. 2). The following description of the data is given by Atkinson (1985, p. 130):

> The air flow [X_1] measures the rate of operation of the plant. The nitric oxides produced are absorbed in a counter-current absorption tower: X_2 is the inlet

temperature of cooling water circulating through coils in this tower and X_3 is proportional to the concentration of acid in the tower. Small values of the response correspond to efficient absorption of the nitric oxides.

The stack loss data have been considered by many authors including Daniel and Wood (1980) and Atkinson (1985). The general consensus is that predictor X_3 (acid concentration) should be dropped from the model and that observations 1, 3, 4, and 21 are outliers. Single deletion diagnostics for all 21 observations for the model with predictors X_1, X_2, and X_3 provide little evidence for the presence of outliers, but robust analyses typically identify these masked outliers.

Below, we consider outlier identification and variable selection for the stack loss data. Transformations are not considered in our analysis; however, there is some evidence that inclusion of a quadratic term (x_1^2) or an interaction term (x_1x_2) will

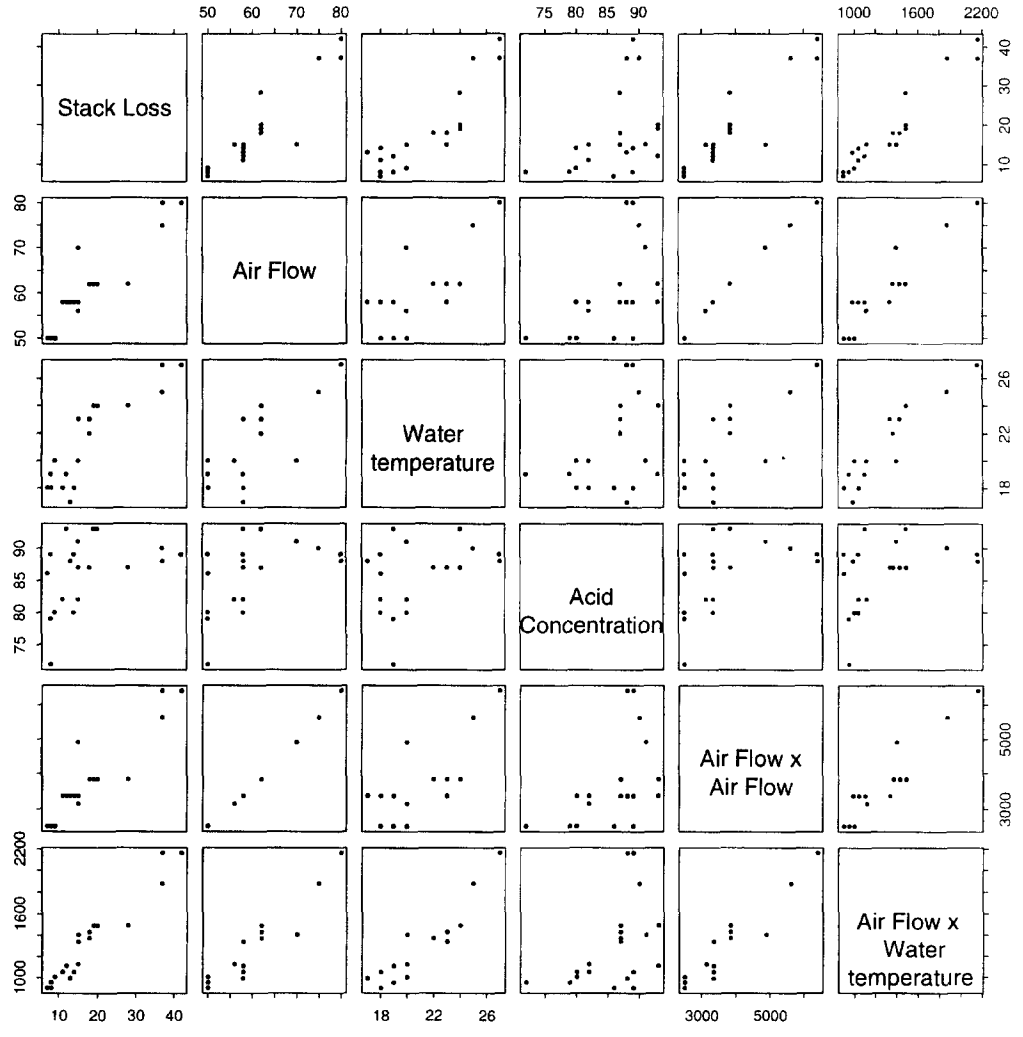

Fig. 2. Scatter plots of stack loss data.

Table 3
Stack loss data: models with the 10 highest posterior model probabilities.

Predictors	Outliers (obs. number)	Posterior model prob. (%)
X_1 X_2	1 3 4 21	23
X_1	4 21	17
X_1 X_2	4 21	9
X_1 X_2	1 3 4 13 21	4
X_1 X_2	21	3
X_1	21	3
X_1 X_2	1 2 3 4 21	3
X_1	1 3 4 21	3
X_1	4 13 21	3
X_1 X_2	3 4 21	2

lead to a better fitting model (Fig. 2). In addition, some authors have suggested that transformation of the response may be appropriate (e.g., Atkinson, 1985; Chambers and Heathcote, 1981). Daniel and Wood (1980) explore the possibility of a temporal relationship in the data. We have chosen not to explore further these issues in this paper.

For these data, the R^2 for the full model is 0.91. As this is a high value of R^2, we again used hyperparameter set 2 for this analysis. We chose $\pi = 0.1$ and $K = 7$.

We used the prescreening method described in Section 3.3 to identify potential outliers. This method indicated that 9 of the 21 observations (observations 1, 2, 3, 4, 8, 13, 14, 20, 21) were potential outliers. Using the set of potential outliers identified by the prescreening procedure, we calculated the posterior model probabilities for all possible combinations of variables and outliers.

The models with the 10 highest posterior model probabilities are shown in Table 3. The model with the highest posterior model probability includes predictors X_1 and X_2, and outliers 1, 3, 4, and 21. Thus, the model with the highest posterior model probability includes the four masked outliers.

The posterior probability that the coefficient for each predictor does not equal 0, i.e. $\Pr(\beta_i \neq 0|D)$, is obtained by summing the posterior probabilities across models containing each predictor. Air flow to the plant, X_1, and cooling water temperature, X_2, both received support from the data with $\Pr(\beta_i \neq 0|D) = 1$ and 0.62, respectively, while acid concentration, X_3, did not with $\Pr(\beta_3 \neq 0|D) = 0.06$.

The marginal posterior distributions for the coefficients of the predictors for the stack loss data are shown in Fig. 3. The posterior distribution for the coefficient for air flow (β_1) is centered away from 0. The posterior β_1 has two modes showing that there is considerable uncertainty about the value of this coefficient. The posterior distribution of the coefficient for water temperature (β_2) is also centered away from 0. This posterior distribution includes a spike at 0 corresponding to $\Pr(\beta_i = 0|D) = 0.38$. The coefficient for acid concentration (β_3) is centered very near 0, with a large spike at 0 corresponding to $\Pr(\beta_i = 0|D) = 0.94$.

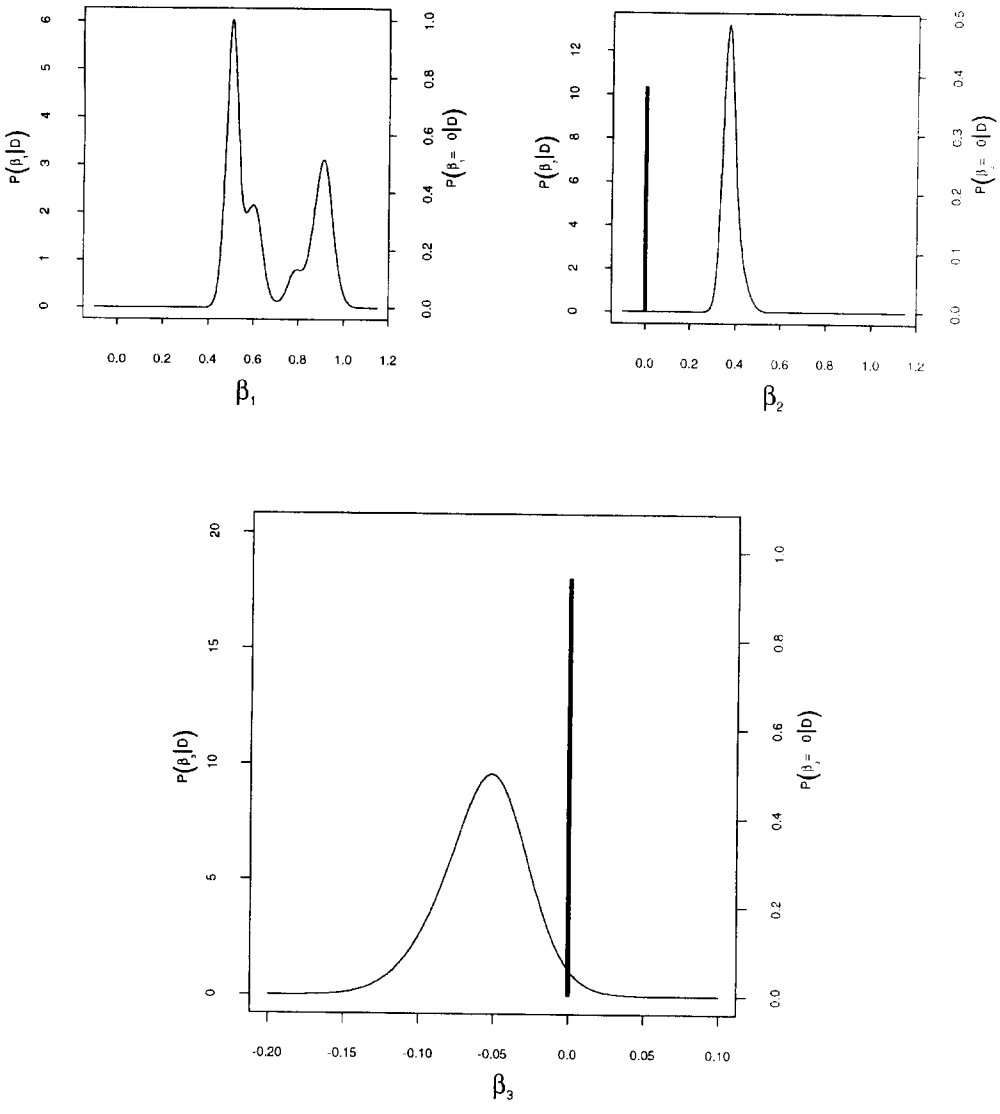

Fig. 3. The marginal posterior density for β_1, β_2, and β_3 of the stack loss data set. The posterior for β_j, $\text{pr}(\beta_j|D)$, is an average across all models. The spike at 0 corresponds to $\text{pr}(\beta_j = 0|D)$. The vertical axis on the left corresponds to the posterior distribution for β_i and the vertical axis on the right corresponds to the posterior distribution for β_i equal to 0.

5. Bayesian model averaging via simultaneous variable selection and outlier identification

A typical approach to data analysis is to carry out a model selection exercise leading to a single "best" model and to then make inferences as if the selected model were the true model. However, this ignores a major component of uncertainty, namely uncertainty about the model itself (Leamer, 1978; Raftery, 1993; Draper, 1995). As a consequence, uncertainty about quantities of interest can be underestimated.

For striking examples of this see Regal and Hook (1991) and Kass and Raftery (1995).

Below we address the problem of model uncertainty and suggest a method for accounting for this uncertainty using Bayesian model averaging. We briefly summarize Bayesian model averaging and describe MC^3. We also discuss several methods for assessing predictive performance. Finally, we provide an example that shows that that model averaging via SVO improves predictive performance as compared to any single model that might reasonably be selected.

5.1. Accounting for model uncertainty via BMA

The standard Bayesian solution to the problem of model uncertainty involves averaging over all possible models. If $\mathcal{M} = \{M_1, \ldots, M_L\}$ denotes the set of all models being considered and if Δ is the quantity of interest such as a future observation or the utility of a course of action, then the posterior distribution of Δ given the data D is

$$\mathrm{pr}(\Delta \mid D) = \sum_{\ell=1}^{L} \mathrm{pr}(\Delta \mid M_\ell, D)\mathrm{pr}(M_\ell \mid D), \tag{2}$$

(Leamer 1978, p.117). This is an average of the posterior distribution under each model weighted by the corresponding posterior model probabilities. We call this "Bayesian model averaging" (BMA). For further details on BMA as applied to linear regression models see Raftery et al. (1994).

Implementation of BMA is difficult for two reasons. First, integrals used to compute $\mathrm{pr}(M_\ell|D)$ can be difficult to solve. Second, the number of terms in (2) can be enormous. Our Bayesian setup described in Section 3 solves the first problem. The MC^3 procedure described below solves the second problem producing an estimate of the Bayesian model average for the entire model space. An alternative method, called Occam's Window, can be used to select models to include in the Bayesian model average (Raftery et al., 1994).

5.2. Markov chain Monte Carlo model composition

For some problems, the number of possible models is very large and it becomes too computationally intensive to compute the posterior probability for every model. To address this problem, we have adapted the Markov chain Monte Carlo model composition (MC^3) approach of Madigan and York (1995) to do BMA over the space of all variables and potential outliers.

Let $\mathcal{M}$ denote the space of models under consideration, including all possible combinations of variables, and all possible combinations of potential outliers. We can construct a Markov chain $\{M(t), t = 1, 2, \ldots\}$ with state space $\mathcal{M}$ and equilibrium distribution $\mathrm{pr}(M_\ell \mid D)$. If we simulate this Markov chain for $t = 1, \ldots, N$, then under certain regularity conditions, for any function $g(M)$ defined on $\mathcal{M}$, the average

$$\frac{1}{N}\sum_{t=1}^{N} g(M(t)) \tag{3}$$

Table 4
Model neighborhood for outliers. The model space includes four predictors (b,c,d,e) and 3 potential outliers (observations 13, 20, 40). The neighborhood of the model with predictors b,c and outliers 13, 20 is given below.

Predictors	Outliers
b	13, 20
c	13, 20
b,c,d	13, 20
$b,c,\ e$	13, 20
b,c	13
b,c	20
b,c	13, 20, 40

converges almost surely to $E(g(M))$ as $N \to \infty$ (Smith and Roberts, 1993). To compute (2) in this fashion set $g(M) = \mathrm{pr}(\Delta \mid M,D)$.

For MC^3, the neighborhood for each model $M \in \mathcal{M}$ is the set of models with either *one predictor or one outlier more* or *one predictor or one outlier less* than the model M itself. For example, if the model space consists of four predictors (b,c,d,e) and 3 potential outliers (observation 13, 20, 40), and the algorithm is currently visiting the model with predictors b,c and outliers 13, 20, then the neighborhood of this model includes the models shown in Table 4.

We define a transition matrix q by setting $q(M \to M') = 0$ for all $M' \notin \mathrm{nbd}(M)$ and $q(M \to M')$ constant for all $M' \in \mathrm{nbd}(M)$. If the chain is currently in state M, we proceed by drawing M' from $q(M \to M')$. It is then accepted with probability

$$\min\left\{1, \frac{\mathrm{pr}(M' \mid D)}{\mathrm{pr}(M \mid D)}\right\}.$$

Otherwise the chain stays in state M.

Software for implementing the MC^3 algorithm is described in the appendix.

5.3. *Assessment of predictive performance*

A primary purpose of statistical analysis is to make forecasts for the future. For MC^3, our specific objective is to compare the quality of the predictions from model averaging with the quality of predictions from any single model that an analyst might reasonably have selected.

To measure performance we randomly split the complete data into two subsets. We run MC^3 using one portion of the data. We call this the training set, D^{T}. We used the remaining portion of the data to assess performance, calling this the prediction set where $D^{\mathrm{P}} = D \setminus D^{\mathrm{T}}$.

The two measures of performance we use are based on the posterior predictive distribution, described below. The first measure of predictive ability is the coverage for 90% prediction intervals. Predictive coverage was measured using the proportion of observations in the performance set that fall in the corresponding 90% posterior prediction interval.

The second measure of predictive ability is the logarithmic scoring rule of Good (1952) where for each event A which occurs, a score of $-\log\{\text{pr}(A)\}$ is assigned. The log predictive score is based on the posterior predictive distribution suggested by Geisser (1980). In this paper, we compare the log predictive score of individual models to the log predictive score of BMA via MC^3. A small log predictive score indicates a model predicts observations in the prediction set well. See Raftery et al. (1994) for details on the computation of predictive coverage and the log predictive score.

We denote the posterior predictive distribution by $\text{pr}(w|\boldsymbol{y})$, where w is an observation in the prediction set, and $\boldsymbol{y}$ is the vector of observations from the training set. To accommodate the possibility that an observation in the prediction set might be an outlier, we adopt a mixture distribution for the posterior predictive distribution.

$$\text{pr}(w \mid \boldsymbol{y}) = \pi p_0(w \mid \boldsymbol{y}) + (1-\pi)p_1(w \mid \boldsymbol{y}), \tag{4}$$

where p_0 is the posterior predictive distribution when w is an outlier and p_1 is the posterior predictive distribution when w is not an outlier.

To assess predictive performance, we incorporate information gleaned from the training set about the prevalence of outliers in the data. To this end, we calculate an updated value of the proportion of outliers, $\hat{\pi} = \sum_{\ell=1}^{L} \text{pr}(M_\ell \mid D)\, q_\ell / n^{\text{T}}$, where L is the number of models, $\text{pr}(M_\ell \mid D)$ is the posterior model probability for model ℓ, q_ℓ is the number of observations identified as outliers under model ℓ, and n^{T} is the number of observations in the training set. To compute the posterior predictive distribution in Eq. (4), we use the maximum of the original value of π used for the training set run, and $\hat{\pi}$. This ensures the possibility of detecting outliers in the prediction set even if there are no outliers in the training data.

Experience to date indicates that for different random splits of the same data set, the algorithms often select different models, but that the log predictive scores for BMA tend to be similar across the random splits.

In the example that follows, we explore how accounting for uncertainty in variable selection and outlier identification influences predictive performance.

5.4. *Example: liver surgery*

A hospital surgical unit interested in predicting survival in patients undergoing a particular type of liver operation collected data on a sample of 108 patients (Neter, Wasserman, and Kutner 1990, henceforth referred to as NWK). Four predictors were extracted from records of the preoperation evaluation of each patient:

X_1 blood clotting score,
X_2 prognostic index, which includes the age of patient,
X_3 enzyme function test score,
X_4 liver function test score.

The response is patient survival time. We used 54 patients for model building and the other 54 patients to assess predictive ability. NWK use the same split of the data for similar purposes. Fig. 4 shows a scatterplot matrix for the entire data.

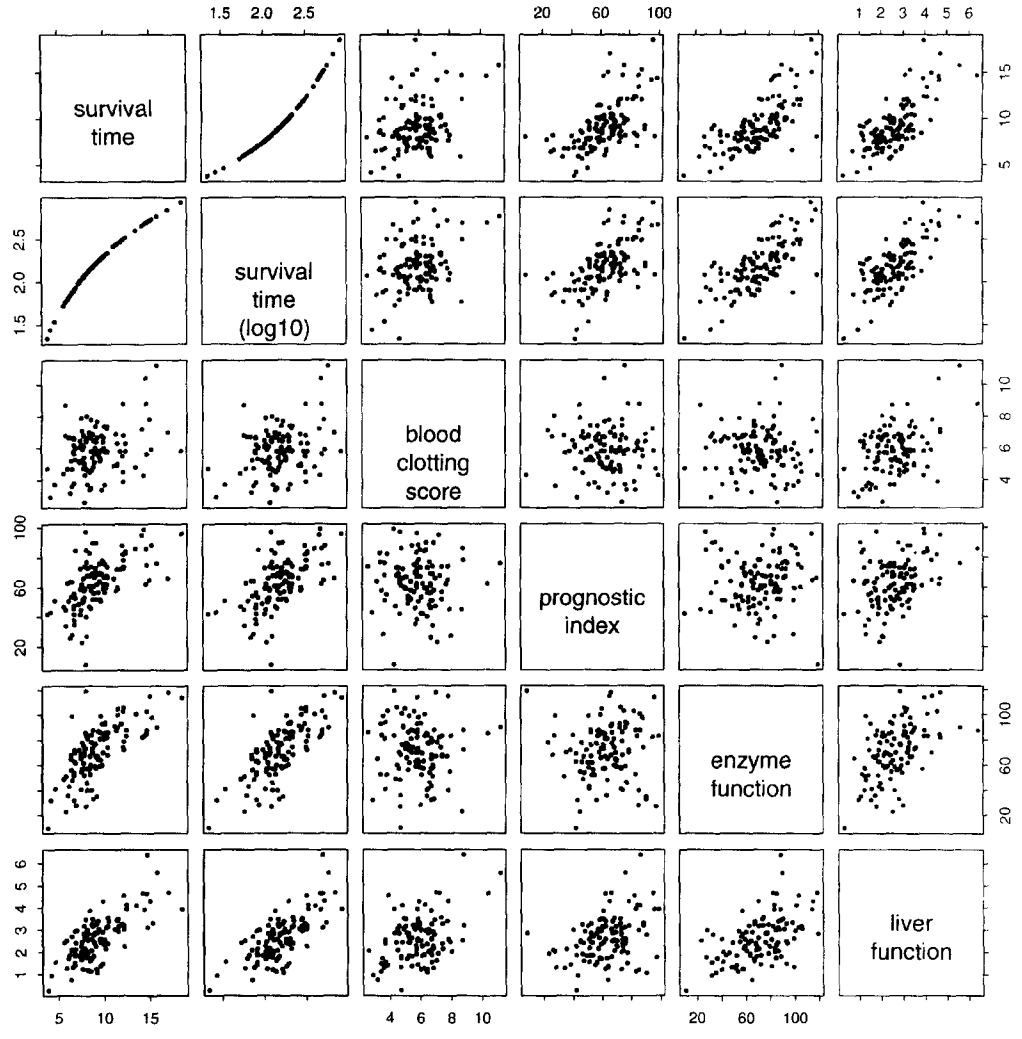

Fig. 4. Scatter plots of the entire liver surgery data without outliers added.

As in NWK we transformed the response logarithmically, so $Y' = \log_{10} Y$. Both forms of the response are shown in Fig. 4. After transforming the response, standard diagnostic checking (e.g., Weisberg, 1985) does not reveal any gross violations of the assumptions underlying normal linear regression.

To demonstrate MC^3 on a data set with known outliers, we introduced artificial outliers in the liver surgery data. To generate outliers we multiplied the first five responses in the training set and the last two responses in the prediction set by 2 before logarithmically transforming the response. We will call these seven observations the "simulated outliers." While multiplying the response by 2 amounts to a shift in mean, we are using a variance-inflation model to accommodate outliers. However, by increasing the values of these 5 observations, we are, in effect, increasing the variance as well. The goal of this exercise is to determine whether our method correctly identifies these aberrant observations.

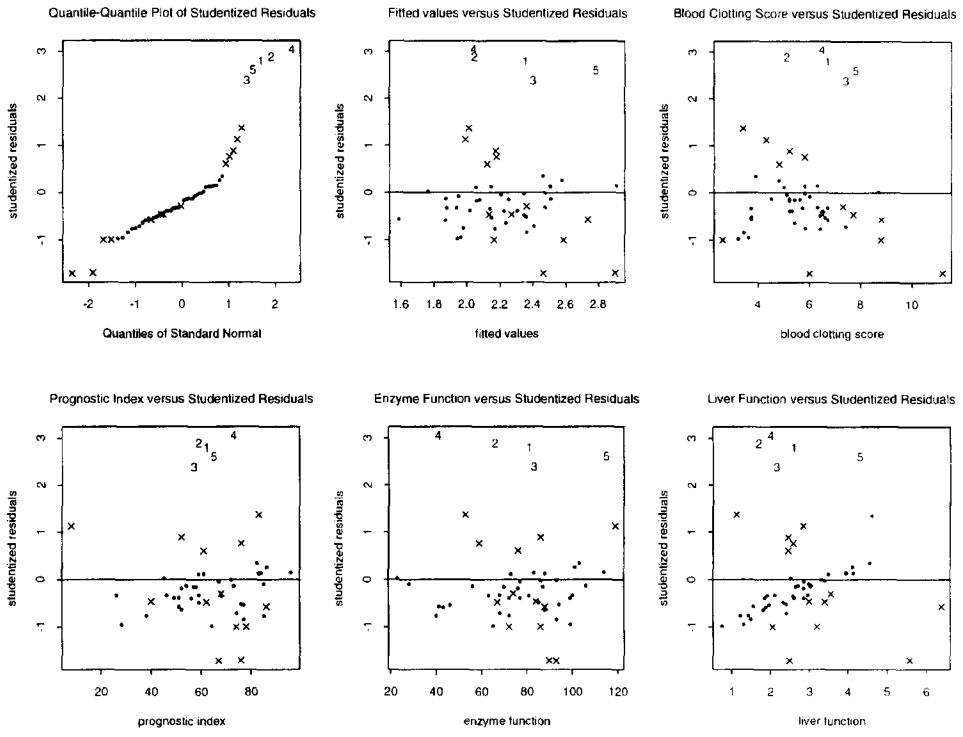

Fig. 5. Residual plots of the liver surgery data for the training set (with outliers added). Observations 1–5 are denoted by numbers and the 13 other observations in the set of potential outliers (observation number 9, 15, 19, 22, 27, 28, 30, 37, 38, 39, 43, 46, 54) are denoted by a × symbol.

For the training set, diagnostic plots of the residuals indicate that the simulated outliers (observations 1–5) are quite different from the rest of the observations (Fig. 5). The absolute values of Studentized residuals for the five simulated outliers range from 2.4 to 3.0 while the Studentized residuals for the rest of the data set range from -1.7 to 1.4. However, Weisberg's outlier test (1985), which is based on the Studentized residuals, does not indicate that these observations are outlying. So while visual inspection of the diagnostic plots might lead to the conclusion that the simulated observations are outlying, there is some uncertainty about whether or not this is the case.

The prescreening procedure identified 18 potential outliers in the training set. They are observations 1, 2, 3, 4, 5, 9, 15, 19, 22, 27, 28, 30, 37, 38, 39, 43, 46, and 54. These observations are denoted by the × symbol in Fig. 5. Note that the simulated outliers were all identified as potential outliers by the prescreening procedure.

Since $R^2 < 0.9$, we used hyperparameter set 1, $K = 7$, and $\pi = 0.02$. All possible models (i.e., all possible combinations of variables and outliers) were assumed to be equally likely a priori. In total, 184 models were visited in 20 000 iterations of MC^3. The models with the 10 highest posterior probabilities are shown in Table 5. The model with the highest posterior model probability includes the predictors X_1, X_2, X_3 and the 5 simulated outliers.

The probabilities that the coefficients for each predictor do not equal 0, $\Pr(\beta_i \neq 0|D)$, are 0.91, 0.91, 0.91, and 0.09, respectively. Thus, there is little support for

Table 5
Liver surgery data with simulated outliers: MC^3. For the log predictive score, $\hat{\pi} = 0.07$. Predictive coverage % is the percentage of observations in the performance set that fall in the 90% prediction interval

Predictors	Outliers (obs. number)	Posterior model prob. (%) MC^3	Log predictive score	Predictive coverage (%)
X_1 X_2 X_3	1 2 3 4 5	50	25.0	80
X_1 X_2 X_3	None	10	14.0	96
X_1 X_2 X_3	1 2 3 4 5 22	5	32.6	72
X_1 X_2 X_3 X_4	1 2 3 4 5	4	24.9	80
X_1 X_2 X_3	1 2 3 4	3	20.9	85
X_1 X_2 X_3 X_4	None	2	15.1	96
X_1 X_2 X_3	4	2	15.2	96
X_1 X_2 X_3	2	2	14.0	96
X_1 X_2 X_3	1	1	13.2	96
X_1 X_2 X_3	1 2 4	1	17.8	91
MC^3 model averaging			19.4	89

Table 6
Liver surgery data with simulated outliers: Outlier posterior probability for each potential outlier, expressed as a percentage. The outlier posterior probability for observations 15, 19, 39, 43, 46, and 54 was approximately equal to 0

	Observation #											
Method	1	2	3	4	5	9	22	27	28	30	37	38
MC^3	78	78	75	79	71	1	6	1	1	1	1	1

inclusion of the predictor liver function test score (X_4) in the model. NWK also conclude that this is not a useful predictor. The outlier posterior probability for each potential outlier is given in Table 6. The simulated outliers (observations 1–5) were identified as outliers in over 70% of the models.

Based on the training set, the estimated value for $\hat{\pi}$ (used in the calculation of prediction coverage and log predictive score for outliers) was 0.07 for MC^3.

Predictive coverage is given in Table 5. The individual models tend to overstate or understate the predictive coverage. Compared to the individual models, model averaging produces more accurate prediction coverage.

Log predictive scores are also given in Table 5. The log predictive score for BMA is smaller than the log predictive score for the model with observations 1–5 identified as outliers (the model with the highest model probability). This indicates that BMA predictively out-performs the model with the highest posterior probability. The model with the highest posterior model probability is also the model which an analyst would probably choose based on visual inspection of standard diagnostic plots of the residuals. The second model in the table has excellent performance according

to the log score; however, this model overstates the predictive coverage by a large amount.

In this example BMA predictively outperforms the model that would be chosen using standard techniques. In addition, our Bayesian framework is successful at simultaneously selecting predictors and identifying masked outliers.

6. Discussion

6.1. Related work

In this work, we adopted a variance–inflation model for outliers. There are other possible formulations including the adoption of heavy-tailed error distributions such as the Student's t-distribution for the outlying observations. This approach to modeling outliers is used by West (1984), Lange et al. (1989), and Besag and Higdon (1993).

Draper (1995) has also addressed the problem of assessing model uncertainty. Draper's approach is based on the idea of *model expansion*, i.e., starting with a single reasonable model chosen by a data-analytic search, expanding model space to include those models which are suggested by context or other considerations, and then averaging over this model class. Draper does not directly address the problem of model uncertainty in variable selection or outlier identification. Clyde et al. (1994) propose a method for model mixing based on a reexpression of the space of models in terms of an orthogonalization of the design matrix. George and McCulloch (1993) developed the stochastic search variable selection (SSVS) method which is similar in spirit to MC^3. In a more recent paper (1994), they suggest an extension of their approach to simultaneous variable selection and outlier identification.

6.2. Conclusions

In this paper we introduced a Bayesian approach to simultaneous variable selection and outlier identification. SVO overcomes the problem that the model you select depends upon the order in which you consider variable selection and outlier identification. We also introduced a method for the identification of multiple outliers which appears to be successful in the identification of masked outliers. Finally, we demonstrated that the model averaging via MC^3 improves predictive performance.

In addition to variable selection and outlier identification, there is also uncertainty involved in the choice of transformations in regression. In Hoeting et al. (1995), we introduce a method to select variables and transformations simultaneously. To broaden the flexibility of the simultaneous approach as well as to improve our ability to account for model uncertainty, we are currently extending our simultaneous approach to include all three components: variable selection, outlier identification, and transformation selection.

Appendix. Software for Implementing MC^3

BMA is a set of S-PLUS© functions which can be obtained free of charge via the World Wide Web address `http://lib.stat.cmu.edu/S/bma` or by sending an e-mail message containing the text "send BMA from S" to the Internet address *statlib@stat.cmu.edu*.

The program MC3.REG performs Markov chain Monte Carlo model composition for linear regression allowing for simultaneous variable selection and outlier identification. The set of programs fully implements the MC^3 algorithm described in Section 5.2.

References

Adams, J.L., A computer experiment to evaluate regression strategies, *Proc. American Statistical Association Section on Statistical Computing* (1991) 55–62.

Atkinson, A.C., *Plots, transformations, and regression* (Clarendon Press, Oxford, 1985).

Atkinson, A.C., Masking unmasked, *Biometrika*, **73** (1986a) 533–541.

Atkinson, A.C., Comments on "Influential Observations, High Leverage Points, and Outliers in Linear Regression," *Statist. Sci.*, **1** (1986b) 397–402.

Atkinson, A.C., Transformations unmasked, *Technometrics*, **30** (1988) 311–318.

Barnett V. and T. Lewis, *Outliers in statistical data*, 3rd ed. (Wiley, New York, 1994).

Besag, J.E. and D.M. Higdon, Bayesian inference for agricultural field experiments, *Bull. Int. Statist. Inst.*, **55** (1993) 121–136.

Blettner, M. and W. Sauerbrei, Influence of model-building strategies on the results of a case-control study, *Statist. Med.*, **12** (1993) 1325–1388.

Bloomfield, P. and W.L. Steiger, *Least absolute deviations: theory, applications, and algorithms*, (Birkhäuser, Boston, 1983).

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

Brownlee, K.A., *Statistical theory and methodology in science and engineering*, 2nd edn. (Wiley, New York, 1965).

Chambers, R.L. and C.R. Heathcote, On the estimation of slope and identification of outliers in linear regression, *Biometrika*, **68** (1981) 21–33.

Chatterjee, S. and A.S. Hadi, Influential observations, high leverage points, and outliers in linear regression, *Statist. Sci.*, **1** (1986) 379–416.

Clyde, M., H. DeSimone and G. Parmigiani, Prediction via orthogonalized model mixing (Institute of Statistics and Decision Sciences, Duke University, 1994)

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

Draper, D., Assessment and propagation of model uncertainty (with discussion), *J. Roy. Statist. Soc.* B, **57** (1995) 45–97.

Garthwaite, P.H. and J.M. Dickey, Elicitation of prior distributions for variable-selection problems in regression, *Ann. Statist.*, **20** (1992) 1697–1719.

Geisser, S., Discussion on Sampling and Bayes' inference in scientific modeling and robustness (by G.E.P. Box), *J. Roy. Statist. Soc.* A, **143** (1980) 416–417.

George, E.I. and R.E. McCulloch, Variable selection via Gibbs sampling, *J. Amer. Statist. Assoc.*, **88** (1993) 881–890.

George, E.I. and R.E. McCulloch, Fast Bayes variable selection, CSS Technical Report 94-01 (University of Texas at Austin, 1994).

Good, I.J., Rational decisions, *J. Roy. Statist. Soc. B*, **14** (1952) 107–114.

Guttman, I., R. Dutter, and P.R. Freeman, Care and handling of univariate outliers in the general linear model to detect spuriousity, *Technometrics*, **20** (1978) 187–193.

Hadi, A.S., A stepwise procedure for identifying multiple outliers in linear regression, *Proc. American Statistical Association Section on Statistical Computing* (1990) 137–142.

Hadi, A.S., A new measure of overall potential influence in linear regression, *Comput. Statist. Data Anal.*, **14** (1992) 1–27.

Hawkins, D.M., *Identification of outliers* (Chapman and Hall, London, 1980).

Heiberger, R. and R.A. Becker, Design of an *S* function for robust regression using iteratively reweighted least squares, AT&T Statistics Research Report (1991).

Hettmansperger, T.P. and S.J. Sheather, A cautionary note on the method of least median squares, *Amer. Statist.*, **46** (1992) 79–83.

Hoaglin, D.C., F. Mosteller and J.W. Tukey (Eds.) *Understanding robust and exploratory data analysis* (Wiley, New York, 1983).

Hoeting, J.A., Accounting for model uncertainty in linear regression, Ph.D. dissertation (University of Washington, 1994) (see http://www.stat.colostate.edu).

Hoeting, J., A.E. Raftery, D. Madigan, Simultaneous Variable and Transformation Selection in Linear Regression, Technical Report 9506 (Department of Statistics, Colorado State University, 1995).

Kadane, J.B., J.M. Dickey, R.L. Winkler, W.S. Smith and S.C. Peters, Interactive elicitation of opinion for a normal linear model, *J. Amer. Statist. Assoc.*, **75** (1980) 845–854.

Kass, R.E. and A.E. Raftery, Bayes factors, *J. Amer. Statist. Assoc.*, **90** (1995) 773–795.

Kianifard, F. and W.H. Swallow, Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression, *Biometrics*, **45** (1989) 571–585.

Lange, K.L., R.J.A. Little and J.M.G. Taylor, Robust statistical modeling using the *t*-distribution, *J. Amer. Statist. Assoc.*, **84** (1989) 881–896.

Leamer, E., *Specification searches: ad hoc inference with nonexperimental data* (Wiley, New York, 1978).

Madigan, D. and A.E. Raftery, Model selection and accounting for model uncertainty in graphical models using Occam's Window, *J. Amer. Statist. Assoc.*, **89** (1994) 1535–1546.

Madigan, D. and J. York, Bayesian graphical models for discrete data, *Internat. Statist. Rev.*, **63** (1995) 215–232.

Marasinghe, M.G., A multistage procedure for detecting several outliers in linear regression, *Technometrics*, **27** (1985) 395–399.

Neter, J., W. Wasserman and M. Kutner, *Applied linear statistical models*, (Irwin, Homewood, IL, 1990).

Pettit, L.I., Bayes factors for outlier models using the device of imaginary observations, *J. Amer. Statist. Assoc.*, **87** (1992) 541–545.

Raftery, A.E., Approximate Bayes factors and accounting for model uncertainty in generalized linear models, Technical Report 255 (Department of Statistics, University of Washington, 1993). (See http://www.stat.washington.edu/tech.reports)

Raftery, A.E., D. Madigan and J. Hoeting, Bayesian model averaging for linear regression models, Technical Report 9412, (Department of Statistics, Colorado State University, 1994) (See http://www.stat.colostate.edu).

Regal, R. and E.B. Hook, The effects of model selection on confidence intervals for the size of a closed population, *Statist. Med.*, **10** (1991) 717–721.

Rousseeuw, P.J., Least median of squares regression, *J. Amer. Statist. Assoc.*, **79** (1984) 871–888.

Rousseeuw, P.J. and A.M. Leroy, *Robust regression and outlier detection* (Wiley, New York, 1987).

Smith, A.F.M. and G.O. Roberts, Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods, *J. Roy. Statist. Soc.* B, **55** (1993) 3–24.

Taplin, R. and A.E. Raftery, Analysis of agricultural field trials in the presence of outliers and fertility jumps, *Biometrics*, **50** (1994) 764–781.

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

Weisberg, S., *Applied linear regression* (Wiley, New York, 1985).

West, M., Outlier models and prior distributions in Bayesian linear regression, *J. Roy. Statist. Soc.* B, **46** (1984) 431–439.