Computational Statistics (1999) 14:293–314

Computational Statistics

© Physica-Verlag 1999

Computations and analysis in robust regression model selection using stochastic complexity[1]

Guoqi Qian

Department of Statistics, La Trobe University
Melbourne, VIC 3083, Australia

Summary

A stochastic complexity approach for model selection in robust linear regression is studied in this paper. Computational aspects and applications of this approach are the focuses of the study. Particularly, we provide both procedures and a package of S language programs for computing the stochastic complexity and for proceeding with the associated model selection. On the other hand, we discuss how a probability distribution on the set of candidate models may be induced by stochastic complexity and how this distribution may be used in diagnosis to measure the likelihood that a candidate model is selected. We also discuss some strategies for model selection when large number of potential explanatory variables are available. Finally, examples and a simulation study are presented for assessing the finite sample performance of our methods.

Keywords: stochastic complexity, minimum description length, model selection, robust regression, Gibbs sampler, Metropolis-Hastings algorithm.

[1] Research supported by the La Trobe University Central Starter Grant No. 09526.

1 Introduction

An essential task in many regression problems is to screen a large number of potential explanatory variables to select one or a few subsets of them which fit best the information contained in the response variable. Another important task is to see how the above screening process is affected by outliers in the data. Namely, the set of selected models should be robust in the sense that they are indifferent to radical change of a small portion of the data or a small change in all of the data. The principle of minimum description length (MDL) and the associated stochastic complexity method, newly developed from the probabilistic theory of information and coding (c.f. Rissanen 1986, 1987, 1989 and 1996), provide a promising approach to model selection in robust regression.

Using this approach, Qian and Künsch (1996) derived a variable selection criterion for robust linear regression. Under this criterion a model, identified by a subset of the explanatory variables, is preferred to another one if relative to the former the stochastic complexity of the data is smaller. Using an optimal two-step coding scheme, the stochastic complexity of the data relative to the underlying regression model was shown to be approximated by the robust fitting error of the model plus the model complexity — a term depending on the robustness and the signal-to-noise ratio of the model, and the weighted magnitude of the explanatory variables. Large sample asymptotic study reveals that the model, which gives the smallest stochastic complexity, almost surely (or with probability 1) coincides with the simplest true model if it exists and can be finitely parameterized. Thus, if there exists a candidate model which gives a significantly smaller stochastic complexity than all the other candidate models, it can be chosen as the best model and be used to make inferences as if it were the simplest true model.

The current paper gives an exposition to computations and some application issues of the stochastic complexity criterion. Specifically, we will address the methods of computing the robust parameter estimates, the weight functions and the criterion function that are involved in the model selection procedure. We will also introduce a package of S language programs we have written for the computations. We will then investigate two complicated but important cases in applications of model selection, namely, that when no clear-cut best model can be selected and when there are too many candidate models so that the exhaustive selection is not computationally feasible. Finally, we will present some examples and a simulation study to illustrate and assess the proposed methods.

2 The Stochastic Complexity Criterion

In this section an introductory description of the stochastic complexity based model selection criterion for robust linear regression is given. This criterion is proposed and studied in Qian and Künsch (1996).

When studying the dependence of a response variable y on a p-dimensional explanatory variable x, a linear model is usually assumed between y and x. Namely, for a sample of independent observations $(x_1^t, y_1), \cdots, (x_n^t, y_n)$ from (x^t, y), we assume

$$y_i = x_i^t \beta + r_i \tag{1}$$

where β is a p-dimensional unknown parameter and r_i is the error with mean 0 conditional on x_i. Provided that the model (1) is valid, information about the indicated dependence can be obtained from a statistical inference of β based on the data. For validity of the model (1), we usually include in (1) all the explanatory variables available in the first consideration in practice, which results in a so-called full model. The validation of the full model usually can be carried out based on the proper subject knowledge. However, if the full model retains many explanatory variables, its statistical inference is typically inefficient and non-informative. Therefore, a variable selection procedure based on solid principles is essential for proceeding with a good regression analysis. With such a procedure, any important explanatory variables should not be missed, while at the same time no superfluous variables can be included in the model.

Of many possible approaches, a particularly attractive one is the stochastic complexity developed by Rissanen. Stochastic complexity measures the goodness of fit of a model by its ability to compress the data. It is formalized by computing the length of an instantaneously decipherable code which is obtained from an optimal two-step coding scheme determined by the employed model. For an employed parametric model, the two-step scheme first encodes the parameter space, then encodes the data for each given parameter value. The shortest code length obtained in such a way is called the stochastic complexity of the data relative to the employed model. According to the minimum description length principle, a model with smaller stochastic complexity is better in extracting the key information of data. Thus, for a model class considered, one would select as the best model the one having the smallest stochastic complexity.

From the work of Rissanen (1996) and Qian and Künsch (1998) it follows that the stochastic complexity relative to a class of parametric probability densities can be expressed as the minus maximum log-likelihood for the data plus a model complexity term determined by the Fisher information and the maximum likelihood estimator (MLE) of the parameter. This result can be directly applied to the regression model (1) if the ordinary least squares method is used, i.e, the error r_i is given a normal distribution. But the parameter estimation and model selection based on least squares can be seriously affected by one or few outliers in the data. Thus, in robust regression, one only assumes r_i to follow some distribution in an infinite dimensional neighbourhood of the normal. In terms of code length this means that the code of the data conditional on a normal distribution is much longer than that conditional on some distribution in its neighbourhood, if there are some outliers in the data. An optimal representation of this neighbourhood is known to be the so-called least favorable distribution (Huber 1964). Therefore, the code can be constructed using the least favorable distribution. With this argument and other ideas underlying the two-step coding scheme, Qian and Künsch (1996) showed that the stochastic complexity of $Y_n = (y_1, \cdots, y_n)^t$ relative to the regression model (1) can be well approximated by

$$SC(Y_n|X_n) = \sum_{i=1}^{n} \rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\hat{\beta})\} + \frac{p}{2}\ln E\rho_c'' + \frac{1}{2}\ln|X_n^t W_n^2 X_n| + \ln\prod_{j=1}^{p}\frac{|\hat{\beta}_j| + n^{-1/4}}{\sigma} \qquad (2)$$

plus terms irrelevant to model selection. Here, $\rho_c(t) = \frac{1}{2}t^2$ for $|t| < c$ and $c|t| - \frac{1}{2}c^2$ for $|t| \geq c$ is the Huber function used to prevent the model selection from being heavily affected by outliers in the data, and $\rho_c''(t) = 1$ for $|t| < c$ and 0 for $|t| \geq c$. The expectation $E\rho_c'' \equiv E\rho_c''(r_0) = (2\Phi(c) - 1)/(2\Phi(c) - 1 + 2c^{-1}\phi(c))$, where Φ and ϕ are respectively the distribution and the density function of standard normal, is taken with respect to the least favorable distribution for r_0. In addition, $X_n = (x_1, \cdots, x_n)^t$, $W_n = diag(w_1, \cdots, w_n)$ with $w_i = w(x_i) \in (0, 1]$ a weight function measuring the outlyingness of x_i, and σ measures the scale of $w(x_i)r_i$. The M-estimator $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)$ is defined by

$$\hat{\beta} = \arg\min_{\gamma} \sum_{i=1}^{n} \rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\gamma)\}, \qquad (3)$$

which belongs to the class of M-estimators considered in Hampel et al. (1986, section 6.3) and is of the type proposed by Hill and Ryan (see Hill 1977). It can be shown that $\hat{\beta}$ is also the MLE relative to the least favorable distribution, which is essential for the stochastic complexity criterion. Since the objective is model selection, those irrelevant terms in the stochastic complexity can be removed. We then can compute (2) for each candidate regression model and choose as the optimal model the one that minimizes (2).

Although obtained from an information and coding theoretic approach, each term in (2) has a clear statistical interpretation. The first term in (2) is the sum of the robustified fitting errors which shows the goodness of robust fit to the observations by the underlying model. It will decrease if additional explanatory variables are included in the model. But the change of the stochastic complexity also depends on other terms in (2) representing the model complexity. The second term gives the cost of using a robust method, which is 0 if $c = +\infty$ and negative otherwise. Note that $c = +\infty$ corresponds to the least squares method which is non-robust against outliers. Thus a robust method is preferred according to the minimum description length principle. The third term gives the weighted magnitude of the explanatory variables and the last one the generalized signal-to-noise ratio. Therefore, the model complexity in (2) is much more comprehensive than that in many other criteria, e.g. AIC, BIC and Mallows' C_p, where it depends essentially on the dimension of the parameter. One can also see that the model complexity in (2) depends on the Fisher information $I_n(\beta) = \sigma^{-2}(E\rho_c^{''})X_n^t W_n^2 X_n$, a quantity measuring how much information Y_n can provide about β in the underlying model (1). It should be pointed out that the last term in (2) is obtained from encoding the parameter space of β truncated to an optimal precision $O(n^{-1/4})$ (see Qian and Künsch (1998) for the detail). Finally, note that the magnitude of each term in (2) is typically $O(n)$, $O(1)$, $O(\ln n)$ and $\sum_j [O(1)I(\beta_j \neq 0) + O(\ln n)I(\beta_j = 0)]$ respectively, where $I(\cdot)$ is the indicator function.

The expression (2) has to be modified to be invariant. Qian and Künsch (1996) proposed the following modification

$$
\begin{aligned}
SC^{'}(Y_n|X_n) = {} & \sum_{i=1}^{n} \rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\hat{\beta})\} + \frac{p}{2}\ln E\rho_c^{''} \\
& + \frac{1}{2}\ln|X_n^t W_n^2 X_n| + \ln\prod_{j=2}^{p}\left(\frac{|\hat{\beta}_j|}{\sigma} + s_{x(j)}^{-1} n^{-1/4}\right), \qquad (4)
\end{aligned}
$$

where $s_{x(j)}^2 = (\sum_{i=1}^n w_i^2)^{-1}\sum_{i=1}^n w_i^2(x_{ij} - \bar{x}_{.j})^2$ and $\bar{x}_{.j} = (\sum_{i=1}^n w_i^2)^{-1}\sum_{i=1}^n w_i^2 x_{ij}$. Assuming that $x_{i1} \equiv 1$, i.e. the regression contains an intercept, and that the p components of x are linearly independent and the weight $w(x)$ is invariant, it can be shown that $SC^{'}(\cdot)$ is invariant under both scale and shift transformations of y and x.

Suppose that the regression model (1) is the full model under consideration, the set of all candidate models can be identified with $\mathcal{A} = \{\alpha :$ any non-empty subset of $\{1, \cdots, p\}\}$ or a subset of $\mathcal{A}$. Each α in $\mathcal{A}$ represents the sub-model of (1): $y = x_\alpha^t\beta_\alpha + r_\alpha$ where x_α and β_α contains those respective components of x and β indexed by α. Based on the approximated stochastic complexity (4), the following model selection procedure is obtained according to the minimum description length principle:

1. For each candidate model $\alpha \in \mathcal{A}$, compute $SC^{'}(Y_n|X_{\alpha n})$, where $X_{\alpha n}$ consists of those columns of X_n indexed by α.

2. Find from $\mathcal{A}$ the best model α^* which minimizes $SC^{'}(Y_n|X_{\alpha n})$ among all candidate models in $\mathcal{A}$. Or alternately, find a subset of $\mathcal{A}$ which have significantly smaller stochastic complexities than other elements of $\mathcal{A}$.

By Theorem 4.2 of Qian and Künsch (1996) it follows that α^* is almost surely (or with probability 1) the simplest model of those in $\mathcal{A}$ which correctly describe the dependence between y and x. Here, the simplest correct model, denoted by α_0, is so called if and only if each component of β_{α_0} is not zero and none of the nonzero components of β are not included in β_{α_0}. If all the explanatory variables are linearly independent, it can be seen that β_{α_0} uniquely exists. In addition, it can be shown that the influence on the selection procedure by both outliers of y and x is bounded if $w(x)^2x^tx$ is bounded. We refer to Qian and Künsch (1996) for more details about these properties. Finally, note that it is always possible in the finite sample situation that α^* is not the (simplest) correct model. Thus, this model uncertainty has to be taken into account in the finite sample model selection. We will discuss this issue in section 5.

3 Computing the Stochastic Complexity

To compute (4), we must be able to compute $\hat{\beta}$ and σ. In addition, we should have a procedure for choosing the weight function $w(x)$ and the tuning parameter c.

Computing the M-estimator $\hat{\beta}$. From (3) it follows that $\hat{\beta}$ is the solution of

$$\sum_{i=1}^{n} \frac{w_i}{\sigma}\psi_c\{\frac{w_i}{\sigma}(y_i - x_i^t\beta)\}x_i = 0, \tag{5}$$

where $\psi_c(t) = \rho'(t) = t$ for $|t| < c$ and $c \cdot sign(t)$ for $|t| \geq c$. Define $u_i = w_i^2 v_i$ with $v_i = \psi_c\{\frac{w_i}{\sigma}(y_i - x_i^t\beta)\}/\{\frac{w_i}{\sigma}(y_i - x_i^t\beta)\}$. The equation (5) is equivalent to

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} u_i(y_i - x_i^t\beta)x_i = 0. \tag{6}$$

It follows from (6) that

$$\hat{\beta} = (\sum_{i=1}^{n} \hat{u}_i x_i x_i^t)^{-1}(\sum_{i=1}^{n} \hat{u}_i y_i x_i). \tag{7}$$

So $\hat{\beta}$ can be computed with a recursive procedure provided that σ, w_i's and c are given. Namely, after getting the k-th step estimate $\beta^{(k)}$, we compute the weights $u_i^{(k)}$'s, then compute a new estimate $\beta^{(k+1)}$ from (7). This process is continued until the difference between two successive computations is negligible. The above procedure is referred to be the iteratively reweighted least squares (IRLS) method.

It can be shown that the IRLS method used here is convergent for any initial value $\beta^{(0)}$ provided that the design matrix X_n has full rank. To see this, note that from (7) it follows that

$$\beta^{(k+1)} = (\sum_{i=1}^{n} u_i^{(k)} x_i x_i^t)^{-1}[\sum_{i=1}^{n} u_i^{(k)}(y_i - x_i^t\beta^{(k)})x_i + \sum_{i=1}^{n} u_i^{(k)} x_i x_i^t \beta^{(k)}]$$

$$= (\frac{1}{\sigma^2}\sum_{i=1}^n u_i^{(k)} x_i x_i^t)^{-1}(\sum_{i=1}^n \frac{w_i}{\sigma}\psi_c\{\frac{w_i}{\sigma}(y_i - x_i^t\beta^{(k)})\}x_i) + \beta^{(k)}.$$

Thus

$$\sum_{i=1}^n \frac{w_i}{\sigma}\psi_c\{\frac{w_i}{\sigma}(y_i - x_i^t\beta^{(k)})\}x_i = (\frac{1}{\sigma^2}\sum_{i=1}^n u_i^{(k)} x_i x_i^t)(\beta^{(k+1)} - \beta^{(k)}). \quad (8)$$

On the other hand, we will prove in the appendix the following inequality

$$\rho_c(t+h) - \rho_c(t) \le h\psi_c(t) + \min\{\frac{1}{2}, \frac{c}{c+|t|}\}h^2 \quad \text{for any real } h \text{ and } t. \quad (9)$$

Denote $t_i^{(k)} = \frac{w_i}{\sigma}(y_i - x_i^t\beta^{(k)})$. It is easy to see that $u_i^{(k)} = w_i^2\psi_c(t_i^{(k)})/t_i^{(k)}$. Now by applying inequality (9) and property (8) we have

$$\sum_{i=1}^n[\rho_c(t_i^{(k+1)}) - \rho_c(t_i^{(k)})] \le (\beta^{(k)} - \beta^{(k+1)})^t \sum_{i=1}^n \frac{w_i}{\sigma}\psi_c(t_i^{(k)})x_i + (\beta^{(k)} - \beta^{(k+1)})^t$$
$$\times\{\sum_{i=1}^n \min\{\frac{1}{2}, \frac{c}{c+|t_i^{(k)}|}\}\frac{w_i^2}{\sigma^2}x_i x_i^t\}(\beta^{(k)} - \beta^{(k+1)}) = -(\beta^{(k+1)} - \beta^{(k)})^t$$
$$\times\{\sum_{i=1}^n \frac{w_i^2}{\sigma^2}(\frac{\psi_c(t_i^{(k)})}{t_i^{(k)}} - \min\{\frac{1}{2}, \frac{c}{c+|t_i^{(k)}|}\})x_i x_i^t\}(\beta^{(k+1)} - \beta^{(k)}). \quad (10)$$

Since $\psi_c(t_i^{(k)})/t_i^{(k)} - \min\{1/2, c(c+|t_i^{(k)}|)^{-1}\} > 0$ for any $t_i^{(k)} \neq 0$, it follows that the right hand side of (10) is always negative if $\beta^{(k+1)} - \beta^{(k)} \neq 0$ and the design matrix X_n has full rank. Therefore, the sequence $\{\sum_{i=1}^n \rho_c(t_i^{(k)})\}$ is decreasing and convergent, and consequently

$$\lim_{k\to\infty}\sum_{i=1}^n \frac{w_i}{\sigma}\psi_c\{\frac{w_i}{\sigma}(y_i - x_i^t\beta^{(k)})\}x_i = 0. \quad (11)$$

From (11) we can see every accumulation point of the sequence $\{\beta^{(k)}\}$ is a solution $\hat{\beta}$ of (3) or (5). Therefore, assuming that $\hat{\beta}$ uniquely exists, which is typically true when n/p is large (Huber 1981, p.178), we have $\lim_{k\to\infty}\beta^{(k)} = \hat{\beta}$. In practice, we usually choose the weighted least squares solution as the initial value $\beta^{(0)}$, which is easy to obtain and often achieves convergence within a few steps.

Computing an estimator of σ. An estimate of σ is needed not only for computing the stochastic complexity (4), but also for computing $\hat{\beta}$ for each candidate model. Since σ is treated as a nuisance parameter in our model selection procedure, it is better and simpler to estimate σ based on the full model when being used in (4) for all the candidate models. This will ensure that the accumulated robust fitting error, i.e. the first term of (4), decreases as additional explanatory variables are included in the model. However, when being used for computing $\hat{\beta}$, it is clear that σ should still be estimated based on the underlying candidate model. Without losing generality we restrict in the following to the case that σ is to be estimated based on the full model. Now, a robust estimate of σ can be obtained by using essentially Huber's proposal 2 (Huber 1981, p.137). Namely, $\hat{\sigma}$ is the solution of the equation

$$\sum_{i=1}^{n} \psi_c^2\{\frac{w_i}{\sigma}(y_i - x_i^t\hat{\beta})\} = (n-p)\gamma(c). \tag{12}$$

where $\gamma(c) = 2\Phi(c) - 1 - 2c\phi(c) + 2c^2(1-\Phi(c))$ is chosen in such a way that it is the expectation of the left hand side of (12) if $w_i(y_i - x_i\beta) = w_i r_i$ has a $\mathcal{N}(0,\sigma)$ distribution. The equation (12) can be solved by the following recursive method.

$$(\sigma^{(k+1)})^2 = \frac{1}{(n-p)\gamma(c)} \sum_{i=1}^{n} \psi_c^2\{\frac{w_i}{\sigma^{(k)}}(y_i - x_i^t\beta^{(k)})\}(\sigma^{(k)})^2. \tag{13}$$

In practice, the estimates $\hat{\beta}$ and $\hat{\sigma}$ are computed by simultaneous iterations from (13) and

$$\beta^{(k+1)} = (\sum_{i=1}^{n} \tilde{u}_i^{(k)} x_i x_i^t)^{-1} (\sum_{i=1}^{n} \tilde{u}_i^{(k)} y_i x_i) \tag{14}$$

with $\tilde{u}_i^{(k)} = w_i^2\psi_c\{\frac{w_i}{\sigma^{(k)}}(y_i - x_i^t\beta^{(k)})\}/\{\frac{w_i}{\sigma^{(k)}}(y_i - x_i^t\beta^{(k)})\}$. The convergence property of this type of simultaneous iterations is studied in detail by Huber (1981, section 7.8) for a more general situation. From his results, it follows that any accumulation point of the sequence $\{\beta^{(k)}, \sigma^{(k)}\}$ minimizes the following function

$$Q(\beta,\sigma) = \sum_{i=1}^{n} \sigma\rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\beta)\} + \frac{1}{2}(n-p)\gamma(c)\sigma$$

in a decreasing manner, which is also a solution of (5) and (12). Instead of using Huber's proposal 2, one could use Hample's median absolute deviation (Hampel 1974, p.388) to estimate σ, i.e. estimate σ by

$$1.4826 \times \text{median}_i\{w_i(y_i - x_i\beta^{(k)})\}$$

at each iteration. Although the empirical evidence of using the latter method is generally good, a rigorous proof of its convergence property seems not available.

Choosing the weight function $w(\cdot)$. Ideally $w(x)$ should be determined by a model which correctly describes the dependence between y and x. But whether a model is correct or not is unknown before proceeding with the model selection. In addition, the penalty of using a wrong model for determining $w(x)$ is not given in the criterion (4). Due to these facts, we suggest that $w(x)$ be determined based on the full model. We believe that the full model is either a correct model if existent or otherwise the one with the most information available. Based on the full model, Qian and Künsch (1996) proposed that

$$w(x) = w_b(x^tBx) \quad \text{where } w_b(t) = \min(1, \tfrac{b}{\sqrt{t}}) \tag{15}$$

with b chosen a priori (e.g. $b = p$) and B a positive definite matrix determined by

$$\frac{2\Phi(c) - 1}{2\Phi(c) - 1 + 2c^{-1}\phi(c)} \frac{1}{n} \sum_{i=1}^{n} w_b(x_i^tBx_i)^2 x_i x_i^t = B^{-1}. \tag{16}$$

By using (15) and (16), the M-estimator $\hat{\beta}$ possesses a robustness property called the bounded self-standardized sensitivity. The expression (15) implies that the influence of x will be weighted down if x^tBx is larger than a given value b. Clearly, the matrix B can be computed with a recursive procedure once b and c are fixed. But this procedure may not converge since the solution B of (16) may not exist or may be multiple. Empirical study shows that the procedure is convergent if b is large enough, but all the weights w_i's equal 1 if b is too large. Finally, note that there are other proposals for computing $w(x)$ in Qian and Künsch (1996). They will not be expounded here.

For the weight function $w(x)$ considered here, one can easily see that it is invariant under a scale transformation. One can also show that it is invariant under a shift transformation provided that the intercept term $x_{i1} \equiv 1$ is true all the time. To see this, let us define a shift transformation $x' = x + a$ where a is a $p \times 1$ vector with the first component a_1 being 0. Denote by B' the counterpart of B under the shift transformation. Knowing that $x_{i1} \equiv 1$, it is easy to verify that $x_i' = (I + (a\ 0))x_i$, $(I + (a\ 0))^{-1} = I - (a\ 0)$ and $B = (I + (a\ 0))^tB'(I + (a\ 0))$. Thus, $w_b(x_i'^tB'x_i') = w_b(x_i^tBx_i)$.

Choosing the tuning parameter c. The smaller the parameter c is, the more robust is the model selection procedure, but at the same time the procedure is also less efficient. We will choose the well-known value 1.345 for c so that $\hat{\beta}$ has efficiency 0.95 when r_i follows a normal distribution. The relationship between c and its corresponding efficiency is discussed in detail in Huber (1981,p.91) and Hampel et al. (1986, p.399).

4 Description of software

The S language (Becker, Chambers and Wilks 1988) provides a very flexible environment for analyzing data. We have written a package of S functions, called msrob, for the robust regression model selection using the stochastic complexity and other related criteria. There are two key functions in the package: xrlm.select and xrlm. The package msrob can be obtained free of charge via the WWW address http://lib.stat.cmu.edu/S/msrob or by sending an e-mail message containing the text "send msrob from S" to statlib@stat.cmu.edu.

The function xrlm.select is used to compute criterion values for all the candidate models and select the optimal model. Its syntax is

```
xrlm.select(data, modset, xweights, sw, k,
                   criterion, xweight.method, ...)
```

The data argument is an $n \times p$ matrix of observations with the first column corresponding to the response and the others the $p - 1$ explanatory variables available. (We assume in addition an intercept term is included in each model considered.) The modset argument, an $m \times (p - 1)$ matrix with either 1 or 0 values, gives a set of m candidate models from $\mathcal{A}$ for selection. Each row of modset specifies a candidate model consisting of those explanatory variables indexed by 1. For example, for $p = 5$ the row $(1, 0, 1, 0)$ identifies the model consisting of x_1, x_3 and the intercept term. The weights w_i's are given by the $1 \times n$ vector xweights. If xweights is not specified, the xweight.method argument tells which method is used to calculate w_i's. For example, xweight.method="implicit.lfn" means the w_i's are computed from (15) and (16). If sw is not given, the scale σ is estimated iteratively by the modified Hampel's median absolute deviation. But at the iteration number given by the sw argument (usually take sw=1 or 2), σ will be estimated by solving (12). The argument k, which is the tuning parameter c in section 3, has the default value 1.345. The criterion argument tells which criterion is used in model selection. It has four possible values: Stochastic-Complexity, Ronchetti-AICR (Ronchetti 1985), Hampel-AICR (Hampel 1983) and Robust -BIC (Machado 1993). There are many other arguments in xrlm.select which are specified by the default values. But one can also change them to achieve more flexibility. The function xrlm.select returns the optimal model selected by each of the four criteria, the criterion values of each candidate model and other useful information.

The function xrlm is used to fit a robust regression model according to (3). Its syntax is

```
xrlm(formula, data, xweights, sw, k, ...)
```

This function adapts the rlm function provided by Venables and Ripley (1994, p.216) to include an xweights argument. Its output includes the value of $\hat{\beta}$, the estimate of σ and many other results. Further, the output can be inherited and analyzed by other S functions to draw statistical inference.

5 Some practical issues

5.1 When there are many explanatory variables.

To implement the model selection procedure proposed in Section 2, one need to compute the stochastic complexity for each of the 2^p candidate models in $\mathcal{A}$. This would involve very intensive computations even if p is moderate (e.g. $p = 20$), and would not be computationally feasible if p is large (e.g. $p = 100$). Various modifications based on classical criteria, such as the stepwise and the leaps and bounds, have been proposed for dealing with this complication (refer to e.g. Miller 1990). Here we will study two new modifications and see how they are adapted to the stochastic complexity criterion. The first one is based on the kick-one-off method proposed in Rao and Wu (1989). The second one is based on the Markov chain Monte Carlo methods such as the Gibbs sampler and the Metropolis algorithm.

Applying the kick-one-off approach to the stochastic complexity criterion, the following procedure can be used for model selection from $\mathcal{A}$. First, compute the stochastic complexity $SC'(Y_n|X_{\alpha(i)n})$ for each model $\alpha(i) = \{1, \cdots, i-1, i+1, \cdots p\}$ $(i = 1, \cdots, p)$ obtained by deleting the i-th component $x_{(i)}$ of the explanatory variable x from the full model $\alpha_f = \{1, \cdots, p\}$. Then compare $SC'(Y_n|X_{\alpha_f n})$ with each $SC'(Y_n|X_{\alpha(i)n})$. Finally, select as the appropriate model α' which consists of those β_i's satisfying $SC'(Y_n|X_{\alpha(i)n}) > SC'(Y_n|X_{\alpha_f n})$.

With the kick-one-off approach one need only compute the stochastic complexity for $p + 1$ candidate models in $\mathcal{A}$, which is computationally feasible even when p is very large. From Theorems 4.1 and 4.2 of Qian and Künsch (1996) we know that, when α_f is a correct model and other weak regularity conditions are satisfied, $SC'(Y_n|X_{\alpha(i)n}) > SC'(Y_n|X_{\alpha_f n})$ with probability 1 if and only if $\beta_i \neq 0$. Thus, the selected model α' is to be the simplest correct model α_0 with probability 1. However, in the finite sample situation it is fairly possible that α' is neither α_0 nor the best model α^* that minimizes SC' over $\mathcal{A}$. In particular, the kick-one-off approach of the stochastic complexity criterion stands a fair chance of missing out an important explanatory variable in the finite sample situation. This can be seen in Example 2 at the end of this section.

Probably a better computation approach for applying the stochastic complexity criterion is to use Markov chain Monte Carlo (MCMC) methods which have been used in a variety of Bayesian model selection procedures (see e.g. George and McCulloch 1997). The basic idea is to simulate by an MCMC method a sample of candidate models from $\mathcal{A}$ so that the marginal sampling distribution converges to the posterior distribution of the model. Then the model with the highest posterior probability can be estimated by the one with the highest frequency in the sample. This approach avoids computing the posterior probabilities for all the 2^p models. Because models with high posterior probabilities are most likely to appear quickly, such models can sometimes be identified in a relatively short simulated sample.

The MCMC methods can also be applied to select models by the non-Bayesian stochastic complexity. This can be seen from the following exposition. First, note that the MCMC methods provide a powerful tool for simulating random samples indirectly from complex or non-standard distributions. By simulating a large enough sample, any characteristic of a distribution and its associated likelihood function can be calculated to the desired degree of accuracy. See, for example, Tanner (1996) for an introduction of the MCMC methods. Next, from the derivation of stochastic complexity (Qian and Künsch 1998) it follows that $SC'(Y_n|X_{\alpha n})$ is an approximate length of an optimal instantaneously decipherable code for the data Y_n relative to a model α. By Kraft inequality (see Section 2.2.1 of Rissanen 1989), $\exp\{-SC'(Y_n|X_{\alpha n})\}$ identifies, up to a constant factor, a predictive probability distribution for Y_n given model α. It can also be regarded as a likelihood function for α if Y_n is given. Thus the stochastic complexity criterion can also be stated as to select the model α^* that maximizes the likelihood $\exp\{-SC'(Y_n|X_{\alpha n})\}$ over $\mathcal{A}$. Denote $f(\alpha) = c_1 \exp\{-SC'(Y_n|X_{\alpha n})\}$, with $c_1 = (\sum_{\alpha\in\mathcal{A}} \exp\{-SC'(Y_n|X_{\alpha n})\})^{-1}$, be the normalized distribution induced by $SC'(Y_n|X_{\alpha n})\}$. Suppose we can generate a random sample of models $\{\tilde{\alpha}_i, i = 1,\cdots,N\}$ from $f(\cdot)$. Then for any $\alpha \in \mathcal{A}$, the sample frequency $f_{\alpha,N} = N^{-1}\sum_{i=1}^{N} I(\tilde{\alpha}_i = \alpha)$ satisfies that

$$\lim_{N\to\infty} f_{\alpha,N} = c_1 \exp\{-SC'(Y_n|X_{\alpha n})\} \quad \text{with probability 1.}$$

Thus one can use $f_{\alpha,N}$ to estimate $f(\alpha)$ with variance $V(f_{\alpha,N}) = N^{-1}f(\alpha)(1-f(\alpha))$. And the model $\tilde{\alpha}^*$ with the highest sample frequency $f_{\tilde{\alpha}^*,N}$ can be used as a consistent estimate of α^*.

Now we give the standard use of two common MCMC methods — Gibbs sampler and Metropolis-Hastings (MH) algorithms — for generating a sample from the distribution $\{f(\alpha), \alpha \in \mathcal{A}\}$. These two methods were proposed in Bayesian model selection context by George and McCulloch (1997) and Madigan and York (1995) respectively. To simplify the presentation, we introduce a one-to-one transformation $\gamma = \gamma(\alpha)$, where $\gamma = (\gamma_1,\cdots,\gamma_p)$ is a $1\times p$ vector with $\gamma_i = 1$ if $i \in \alpha$ and 0 if $i \notin \alpha$, $i = 1,\cdots,p$. Denote $\mathcal{A}_1 = \gamma(\mathcal{A})$, and $\gamma_{(i)} = (\gamma_1,\cdots,\gamma_{i-1},\gamma_{i+1},\gamma_p)$. Now simulating a sample from $\{f(\alpha), \alpha \in \mathcal{A}\}$ is equivalent to simulating a sample from $\{f(\gamma) = f(\alpha), \gamma \in \mathcal{A}_1\}$. To be able to use the Gibbs sampler, one need only the conditional distributions $\{f(\gamma_i|\gamma_{(i)}), \gamma_i = 0, 1\}$ $(i = 1,\cdots,p)$. Note that

$$\begin{aligned} f(\gamma_i|\gamma_{(i)}) &= \frac{f(\gamma)}{f(\gamma)|_{\gamma_i=0} + f(\gamma)|_{\gamma_i=1}} \\ &= \frac{\exp\{-SC'(Y_n|X_{\alpha n})\}}{\exp\{-SC'(Y_n|X_{\alpha(\gamma)n})\}|_{\gamma_i=0} + \exp\{-SC'(Y_n|X_{\alpha(\gamma)n})\}|_{\gamma_i=1}} \end{aligned}$$

is just a Bernoulli distribution which can be computed easily from (4). Now we give the following Gibbs sampling algorithm which was originally proposed in Bayesian model selection context by George and McCulloch (1997).

- Arbitrarily choose a starting model $\gamma^{(0)} = (\gamma_1^{(0)}, \cdots, \gamma_p^{(0)})$.
- Repeat $j = 1, 2, \cdots, N$.
- The model $\gamma^{(j)} = (\gamma_1^{(j)}, \cdots, \gamma_p^{(j)})$ is obtained by generating $\gamma_i^{(j)}$ from the Bernoulli distribution with the probability

$$f(\gamma_i | \gamma_1^{(j)}, \cdots, \gamma_{i-1}^{(j)}, \gamma_{i+1}^{(j-1)}, \cdots, \gamma_p^{(j-1)}), \quad i = 1, \cdots, p.$$

- Return the model sequence $\{\gamma^{(1)}, \gamma^{(2)}, \cdots, \gamma^{(N)}\}$.

We will discuss the sampling properties of the above sequence together with the one generated from the MH algorithms. To construct an MH algorithm, we need a candidate-generating distribution, denoted by $q(\gamma', \gamma)$, which is a probability distribution of γ for a given γ'. Then the MH algorithm can be summarized as follows for an arbitrary starting value $\gamma^{(0)}$.

- Repeat $j = 1, 2, \cdots, N$.
- Generate a candidate model $\tilde{\gamma}$ from the distribution $q(\gamma^{(j-1)}, \gamma)$ for γ.
- Set $\gamma^{(j)} = \tilde{\gamma}$ with probability

$$\alpha^{MH}(\gamma^{(j-1)}, \tilde{\gamma}) = \min\{\frac{q(\tilde{\gamma}, \gamma^{(j-1)}) f(\tilde{\gamma})}{q(\gamma^{(j-1)}, \tilde{\gamma}) f(\gamma^{(j-1)})}, 1\}. \qquad (17)$$

- Otherwise, set $\gamma^{(j)} = \gamma^{(j-1)}$.
- Return the model sequence $\{\gamma^{(1)}, \gamma^{(2)}, \cdots, \gamma^{(N)}\}$.

If $q(\gamma', \gamma)$ is symmetric in (γ', γ), the MH algorithm turns out to be the Metropolis algorithm with the probability (17) simplifying to

$$\alpha^M(\gamma^{(j-1)}, \tilde{\gamma}) = \min\{\frac{f(\tilde{\gamma})}{f(\gamma^{(j-1)})}, 1\} = \min\{e^{SC'(Y_n | X_{\alpha(\gamma^{(j-1)})n}) - SC'(Y_n | X_{\alpha(\tilde{\gamma})n})}, 1\}.$$

When choosing

$$q(\gamma', \gamma) = \frac{1}{p}, \quad \text{if } \sum_{i=1}^p |\gamma_i - \gamma_i'| = 1,$$

one gets the following Metropolis algorithm proposed by Madigan and York (1995) for Bayesian model selection:

- Repeat $j = 1, 2, \cdots, N$.
- Generate a candidate model $\tilde{\gamma}$ by randomly changing one component of $\gamma^{(j-1)}$.
- Generate a number u from $Uniform(0, 1)$.
- Set $\gamma^{(j)} = \tilde{\gamma}$ if $u \leq \alpha^M(\gamma^{(j-1)}, \tilde{\gamma})$; otherwise set $\gamma^{(j)} = \gamma^{(j-1)}$.
- Return the model sequence $\{\gamma^{(1)}, \gamma^{(2)}, \cdots, \gamma^{(N)}\}$.

Note that in generating $\{\gamma^{(1)}, \gamma^{(2)}, \cdots, \gamma^{(N)}\}$ by the Gibbs sampler, one need compute $SC'(Y_n|X_{\alpha n})$ for Np models in $\mathcal{A}$; while by the Metropolis algorithm one need N such computations.

The sequence $\{\gamma^{(1)}, \gamma^{(2)}, \cdots, \gamma^{(N)}\}$ generated by each of the Gibbs sampler and the MH algorithm is actually a Markov chain but not an independent sequence. Under mild regularity conditions it can be shown that the sequence converges in distribution to an invariant distribution which is just $\{f(\gamma), \gamma \in \mathcal{A}_1\}$. (See, e.g. Smith and Roberts 1993.) Therefore, by exploiting this sequence, any characteristic of $\{f(\gamma), \gamma \in \mathcal{A}_1\}$ can be calculated to a desired accuracy if N is large enough. In practical model selection, we usually discard the first m values of the sequence because their marginal distributions are not close enough to $\{f(\gamma), \gamma \in \mathcal{A}_1\}$. The selection of m depends on the convergence rate of the method, so it can be quite different for the Gibbs sampler and the Metropolis algorithm. The problem of how to select m and N for an MCMC method is still not well settled — except for some empirical studies. For more details we refer to Tanner (1996, chapter 6) and the references listed therein.

5.2 When there is no clear-cut single best model.

The asymptotic theory in Qian and Künsch (1996) asserts that the model α^* converges to the simplest correct model α_0 with probability 1. But in the finite sample situation it is still possible that $\alpha^* \neq \alpha_0$. On the other hand, it is found in many practical cases that there are several candidate models with the SC' values quite close to the smallest and there is no single model with its SC' value significantly smaller than the others. It seems that to select a small group of models as the best is better than to select a single best since the former takes into account the uncertainty in model selection. Here we will base the stochastic complexity to identify a small group of best models.

Since $\exp\{-SC'(Y_n|X_{\alpha n})\}$ has an interpretation as the likelihood that Y_n can be predicted using model α, from discussions in section 5.1, we define a $\kappa \times 100\%$ $(0 < \kappa < 1)$ likelihood set for the simplest correct model α_0 to be

$$\begin{aligned}\mathcal{A}(\kappa) &= \{\alpha : \alpha \in \mathcal{A}, \quad \frac{\exp\{-SC'(Y_n|X_{\alpha n})\}}{\exp\{-SC'(Y_n|X_{\alpha^* n})\}} > \kappa\} \\ &= \{\alpha : \alpha \in \mathcal{A}, \quad SC'(Y_n|X_{\alpha n}) - SC'(Y_n|X_{\alpha^* n}) < -\ln\kappa\}\end{aligned}$$

or equivalently

$$\mathcal{A}_1(\kappa) = \{\gamma : \gamma \in \mathcal{A}_1, \quad \frac{f(\gamma)}{f(\gamma(\alpha^*))} > \kappa\}.$$

A $\kappa \times 100\%$ likelihood set provides not only the best model α^* but also the set of models whose associated SC' values are not greater than the smallest SC' by $-\ln k$. For $\kappa = 0.9, 0.5$, and 0.1, $-\ln\kappa \approx 0.105, 0.693$ and 2.303

respectively. In general, there is no strict rule for choosing κ. Usually we would like κ to be so selected that models in $\mathcal{A}(\kappa)$ have significantly smaller SC' values than the other models.

In practice we regard $\mathcal{A}(\kappa)$ (or $\mathcal{A}_1(\kappa)$) as the set of the best models by the stochastic complexity criterion. Then we can calculate the probability $p(\kappa) = \sum_{\alpha \in \mathcal{A}(\kappa)} f(\alpha)$ and interpret it as the cumulative likelihood that Y_n can be predicted using the models in $\mathcal{A}(\kappa)$ (or $\mathcal{A}_1(\kappa)$). It can be seen that $p(\kappa)$ is a decreasing function of κ.

If there are many explanatory variables in the full model, it is computationally not feasible to find $\mathcal{A}(\kappa)$ and $p(\kappa)$ directly. But as discussed in section 5.1, an MCMC method can be used to provide approximations for $\mathcal{A}(\kappa)$ and $p(\kappa)$. Namely, approximate $\mathcal{A}(\kappa)$ by $\mathcal{A}_N(\kappa) = \{\alpha : \alpha \in \mathcal{A}_N, f_{\alpha,N}/\max f_{\alpha,N} > \kappa\}$ for a sample of generated models $\mathcal{A}_N = \{\tilde{\alpha}_i, i = 1, \cdots, N\}$ and $p(\kappa)$ by $p_N(\kappa) = \sum_{\alpha \in \mathcal{A}_N(\kappa)} f_{\alpha,N}$. We give some examples in the following.

5.3 Examples

Example 1. *The triathlon athletes data* were taken from Kohrt et al. (1987) who studied the performance of a group of 65 male athletes in half-triathlon event over a 6-week period. The data can also be found in Glantz and Slinker (1990, pp.647-648). There are 10 variables in the data: half-triathlon performance time (t min.), age (A years), weight (W kg.), experience (E years), amount of training running (T_R km/week), biking (T_B km/week), and swimming (T_S km/week), and maximum oxygen consumption while running (V_R mL/min/kg), biking (V_B mL/min/kg), and swimming (V_S mL/min/kg).

The objective of the study is to see which variables determine best the athletes' final time when they compete in the triathlon. This was addressed by conducting a variable selection on the full regression model

$$t = \beta_0+\beta_1 A+\beta_2 W+\beta_3 E+\beta_4 T_R+\beta_5 T_B+\beta_6 T_S+\beta_7 V_R+\beta_8 V_B+\beta_9 V_S+r. \quad (18)$$

We applied to variable selection the stochastic complexity criterion. There were in total $2^9 = 512$ sub-models for selection if only considering those including an intercept term. Table 1 gives the 8 best sub-models and their associated SC' values of (4), which were obtained by the exhaustive selection. It is easy to verify that the 8 models, for which the SC' values are not larger than 40.11 ($\approx$ 40), comprise the 17.2% likelihood set with the cumulative likelihood 0.553 for the simplest correct model. From Table 1 we also see the model $A + E + T_R + T_B + V_R$ is a clear-cut best model with its SC' nearly 1 smaller than that of the second best model. Table 2 gives the performance of the kick-one-off approach. It shows that in this example the model selected by the kick-one-off approach is also the best by the exhaustive selection.

To evaluate the performance of the MCMC approach, we first generated by the Gibbs sampler a sample of 1100 models from the 512 sub-models.

It seemed that by discarding the first $m = 100$ models of the sample we could ensure that the marginal sampling distribution was close enough to $\{f(\alpha), \alpha \in \mathcal{A}\}$. It was found that the remaining sample of 1000 models contained 81 distinct models, the sample frequencies(1000 × $f_{\alpha N}$) of which were summarized by the frequency table given in Table 3. (For example, the first column values of Table 3 can be interpreted as there were 57 models each of which had its frequency less than 0.01, or, appeared less than 10 times in the sample.) It was also found that the model $A+E+T_R+T_B+V_R$, the best under the exhaustive selection, had the highest frequency 0.185 (or appeared 185 times) in the sample. The 8 most frequently appeared models in the sample and some of their key information are given in Table 4. It can be seen that Table 4 and Table 1 share 7 common models. The 8 models in Table 4

Table 1. Eight best models from the exhaustive search in Example 1.

Model	SC'
$A+E+T_R+T_B+V_R$	38.35
$A+E+T_R+T_B+V_R+V_B$	39.29
$A+E+T_B+V_R$	39.31
$A+E+T_R+T_B+T_S+V_R$	39.72
$A+E+T_R+T_B+V_R+V_S$	39.73
$A+E+T_S+V_R+V_B$	39.88
$A+W+E+T_R+T_B+V_R$	39.97
$A+E+T_B+V_R+V_B$	40.11

Table 2. SC's for the full and the delete-1-term models in Example 1.

Model	SC'
full model	42.95
A deleted	**44.87**
W deleted	41.47
E deleted	**49.95**
T_R deleted	**43.09**
T_B deleted	**43.85**
T_S deleted	42.19
V_R deleted	**47.58**
V_B deleted	42.59
V_S deleted	41.78

Table 3. Summary of the frequencies of the Gibbs sample in Example 1.

Frequency range	< 10	[10,20)	[20,30)	[30, 40)	≥ 40	Total
# of models	57	11	4	2	7	81

Table 4. The 8 best models selected by the Gibbs sample in Example 1.

Model	SC'	$f(\alpha)$	$f_{\alpha,N}$
$A+E+T_R+T_B+V_R$	38.35	0.1930	0.185
$A+E+T_B+V_R$	39.31	0.0741	0.078
$A+E+T_R+T_B+V_R+V_B$	39.29	0.0754	0.076
$A+E+T_S+V_R+V_B$	39.88	0.0417	0.045
$A+W+E+T_R+T_B+V_R$	39.97	0.0381	0.045
$E+T_S+V_R+V_B$	40.27	0.0282	0.041
$A+E+T_R+T_B+T_S+V_R$	39.72	0.0488	0.040
$A+E+T_R+T_B+V_R+V_S$	39.73	0.0484	0.038
Total		0.5477	0.548

comprise a set of sub-models with the cumulative likelihood 0.5476. The total frequency of these 8 models is 0.548, which is a very precise approximation for 0.5476.

We also evaluated the performance of the Metropolis algorithm by generating a sample of 2000 models with the first $m = 1000$ being discarded. The results are presented in Tables 5 and 6. It cane be seen the performance of the Metropolis algorithm is a little bit worse than that of the Gibbs sampler in our example, which may be partially justified by the fact that the Gibbs sampler involves more computations for generating a sample of a fixed size.

Table 5. Summary of frequencies of the Metropolis sample in Example 1.

Frequency range	< 10	[10,20)	[20,30)	[30, 40)	≥ 40	Total
# of models	37	9	5	4	7	62

Table 6. The 8 best models chosen by the Metropolis sample in Example 1.

Model	SC'	$f(\alpha)$	$f_{\alpha,N}$
$A + E + T_R + T_B + V_R$	38.35	0.1930	0.169
$A + E + T_S + V_R + V_B$	39.88	0.0417	0.081
$A + E + T_R + T_B + V_R + V_B$	39.29	0.0754	0.051
$A + E + T_R + T_B + V_R + V_S$	39.73	0.0484	0.049
$E + T_S + V_R + V_B$	40.27	0.0282	0.048
$A + E + T_S + V_R$	40.56	0.0212	0.047
$A + E + T_R + T_B + T_S + V_R$	39.72	0.0488	0.044
$A + E + T_B + V_R$	39.31	0.0741	0.037
Total		0.5308	0.526

Example 2. *The Hoffstedt highway data*, taken from Weisberg(1985, p.206), relate the automobile accident rate per million vehicle miles (Y) to 13 potential explanatory variables ($X_1, \cdots, X_9$ and $X_a, \cdots, X_d$). This data set has been used by Weisberg (1985, pp. 205-221) to illustrate variable selection and by Ronchetti and Staudte (1994) to illustrate robust C_p method.

If only considering those models with an intercept term, there are in total $2^{13} = 8192$ possible sub-models for selection. We have applied the stochastic complexity criterion for the variable selection. Four different approaches — the exhaustive, the kick-one-off, the Gibbs sampler and the Metropolis algorithm — were used with settings the same as those in Example 1. The results are presented in Tables 7 to 12, which correspond to Tables 1 to 6 for Example 1. We see here the kick-one-off approach selected the model with the only variable X_4, which is not the same as the one selected by the exhaustive approach. The two MCMC approaches both identified the model $X_1 + X_4 + X_9$ with the highest frequency. But their performances are worse than in Example 1, in the sense that the 8 best models selected by the MCMC approaches differ more from those by the exhaustive search. On the other hand we see the model $X_1 + X_4 + X_9$, with the smallest SC' value 20.58, is not the clear-cut best model since at least another model $X_1 + X_9$ has a very close SC' value 20.62. This may justify the selection of a small set of sub-models as the most appropriate models.

Table 7. Eight best models from the exhaustive search in Example 2.

Model	SC'
$X_1 + X_4 + X_9$	20.58
$X_1 + X_9$	20.62
$X_3 + X_4 + X_9$	21.33
$X_3 + X_9$	21.38
X_9	21.40
$X_1 + X_8 + X_9$	21.48
$X_1 + X_4 + X_8 + X_9$	21.50
$X_4 + X_9$	21.56

Table 8. SC's for the full and the delete-1-term models in Example 2.

Model	SC'
full model	31.77
X_1 deleted	31.47
X_2 deleted	31.15
X_3 deleted	30.65
X_4 deleted	**31.85**
X_5 deleted	30.35
X_6 deleted	30.90
X_7 deleted	30.96
X_8 deleted	31.18
X_9 deleted	31.75
X_a deleted	30.65
X_b deleted	31.00
X_c deleted	31.24
X_d deleted	30.78

Table 9. Summary of the frequencies of the Gibbs sample in Example 2.

Frequency range	< 10	[10,20)	≥ 20	Total
# of models	566	4	2	572

Table 10. The 8 best models selected by the Gibbs sample in Example 2.

Model	SC'	$f(\alpha)$	$f_{\alpha,N}$
$X_1 + X_4 + X_9$	20.58	0.0227	0.026
$X_1 + X_9$	20.62	0.0219	0.025
$X_1 + X_8 + X_9$	21.48	0.0092	0.016
$X_3 + X_9$	21.38	0.0102	0.013
$X_3 + X_4 + X_9$	21.33	0.0107	0.012
$X_1 + X_9 + X_a$	22.20	0.0045	0.010
X_9	21.40	0.0099	0.009
$X_1 + X_2 + X_4 + X_9$	22.04	0.0053	0.009
Total		0.0944	0.120

Table 11. Summary of frequencies of the Metropolis sample in Example 2.

Frequency range	< 10	[10,20)	≥ 20	Total
# of models	376	6	2	384

Table 12. The best models chosen by the Metropolis sample in Example 2.

Model	SC	$f(\alpha)$	$f_{\alpha,N}$
$X_1 + X_4 + X_9$	20.58	0.0227	0.024
$X_1 + X_4 + X_8$	22.63	0.0029	0.022
$X_1 + X_4 + X_8 + X_c$	22.05	0.0052	0.017
$X_1 + X_4 + X_9 + X_d$	22.17	0.0046	0.012
$X_1 + X_3 + X_4 + X_9$	21.60	0.0082	0.011
$X_1 + X_3 + X_4 + X_9 + X_c$	23.28	0.0015	0.011
$X_1 + X_2 + X_4 + X_9$	23.56	0.0011	0.011
$X_1 + X_9$	20.62	0.0219	0.010
Total		0.0681	0.118

6 A Simulation Study

We carried out a simulation study to evaluate the robustness of our stochastic complexity criterion as well as three other criteria — the two versions of robust AIC by Ronchetti(1985) and Hampel(1983) and the robust BIC by Machado (1993). We considered $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + r$ as the full model. So there were in total $2^6 = 64$ possible sub-models with an intercept term. The sample size n was 30. The variables X_1 to X_6 were generated independently from Uniform$[0,1]$ except that the first observation of each X_i was 3 and the second was 5. Thus, the first two sample points were leverage points and had large influence on the regression procedure. Six distributions for the error r were chosen to represent various deviation from normality. They are Normal(0,1), $t_{(3)}$ with 3 degrees of freedom, Cauchy ($t_{(1)}$), log-normal with mean 0 and scale 1 which is asymmetric, slash which is a Normal(0,1) divided by a Uniform$[0,1]$, and ε-normal $0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(0,3)$. The observations of Y were obtained from

$$Y = 1 + 2.5X_1 + 3X_2 - 3X_3 + r \qquad (19)$$

with r generated from one of the six error distributions. The coefficient values were so selected that they would give t-values of about 4 if r were normally distributed. It is clear that the model (19) is the true model. But other models containing X_1, X_2 and X_3 are also correct models[2]. We carried out 200 simulation runs. Table 13 gives the frequencies of selecting the three types—true, other correct and incorrect—of models by each of the four criteria.

From Table 13 we see all the four criteria perform quite well even when the error distribution has a considerable deviation from normality (i.e. $t_{(3)}$, log-

[2] Strictly saying, just specifying which X_i's are included determines only a regression model class with the associated β_i's to be given. A regression model class determined by those including X_1, X_2 and X_3 is a correct class because it contains the true model (19). Since β is mostly unknown and is uniquely estimated by the underlying regression procedure, such a correct model class may be called a correct model for purpose of conciseness.

normal and ε-normal). The relative frequencies of selecting the true model is between 55.5% and 78% for these three error distributions. (Compare with 59% and 84% for the normal error.) They are between 4% and 23.5% in selecting the incorrect models. But when the error distribution is Cauchy or slash, neither of the criteria works well in selecting the correct models. This is probably because Cauchy and slash deviate so much from normal that their population expectations do not exist. Thus a more robust and efficient procedure would be required for this situation. When comparing these four criteria with each other, we see the AIC methods usually have lower frequencies of selecting the true model but higher frequencies of selecting other superfluous correct models than the other two criteria. The stochastic complexity criterion may have little lower frequencies of selecting the true model than the BIC method, but it also has lower frequencies of selecting the incorrect models so has a more stable performance.

Table 13. Frequencies of Different Models Selected in 200 Simulations

	Error Distribution					
Model Category	$\mathcal{N}(0,1)$	$t_{(3)}$	Cauchy	Log-N(0,1)	Slash	ε-N
	Stochastic Complexity Criterion					
True	143	117	30	129	7	135
Other correct	55	49	9	44	5	50
Incorrect	2	34	161	27	188	15
	Ronchettis's Robust AIC					
True	118	111	42	122	20	122
Other correct	81	72	17	64	8	70
Incorrect	1	17	141	14	172	8
	Hampel's Robust AIC					
True	126	116	40	127	17	129
Other correct	72	64	16	55	7	63
Incorrect	2	20	144	18	176	8
	Machado's Robust BIC					
True	168	134	28	151	6	156
Other correct	27	19	7	15	2	27
Incorrect	5	47	165	34	192	17

Appendix. Proof of Inequality (9)

Define $F(h) = \rho_c(t+h) - \rho_c(t) - h\psi_c(t) - \min\{1/2, c(c+|t|)^{-1}\}h^2$ for any given t. Straightforward calculations give the following expressions:
When $t \geq c$

$$F(h) = \begin{cases} -2c(t+h) - c(c+t)^{-1}h^2, & h \leq -c-t \\ \frac{1}{2}(t+h-c)^2 - c(c+t)^{-1}h^2, & -c-t < h < c-t \\ -c(c+t)^{-1}h^2, & h \geq c-t \end{cases}$$

When $t \leq -c$

$$F(h) = \begin{cases} -c(c-t)^{-1}h^2, & h \leq -c-t \\ \frac{1}{2}(t+h+c)^2 - c(c-t)^{-1}h^2, & -c-t < h < c-t \\ 2c(t+h) - c(c-t)^{-1}h^2, & h \geq c-t \end{cases}$$

When $|t| < c$

$$F(h) = \begin{cases} -\frac{1}{2}(t+h+c)^2, & h \leq -c-t \\ 0, & -c-t < h < c-t \\ -\frac{1}{2}(t+h-c)^2, & h \geq c-t \end{cases}$$

It is easy to show that each of the above expressions is not great than 0. For example, given that $t \geq c$ and $h \leq -c-t$, we have $F(-c-t) = c^2 - ct \leq 0$ and $F'(h) = -2c(c+t+h)(c+t)^{-1} \geq 0$, thus $F(h) \leq 0$. Therefore, the assertion of (9) follows.

Acknowledgment: I am grateful to an anonymous referee for the useful comments on the first version of the paper.

References

Becker, R., Chambers, J.M. & Wilks, A. (1988), *The New S language*, Wadsworth, Belmont CA.

George, E.I. & McCulloch, R.E. (1997), 'Approaches for Bayesian variable selection', *Statistica Sinica* **7**, 339-373.

Glantz, S.A. & Slinker, B.K. (1990), *Primer of Applied Regression and Analysis of Variance*, McGraw-Hill, Inc., New York.

Hampel, F.R. (1974), 'The influence curve and its role in robust estimation', *J. Amer. Statist. Assoc.* **69**, 383-393.

Hampel, F.R. (1983), 'Some aspects of model choice in robust statistics', *Proceedings of the 44th Session of ISI, Book 2*, Madrid, 767-771.

Hampel, F.R., Ronchetti, E. M.,Rousseeuw, P. J. & Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

Hill, R.W. (1977), *Robust regression when there are outliers in the carriers*, Ph.D. thesis, Harvard University, Cambridge, Mass..

Huber, P.J. (1964), 'Robust estimation of a location parameter', *Ann. Math. Stat.* **35**, 73-101.

Huber, P.J. (1981), *Robust Statistics*, Wiley, New York.

Kohrt, W.M., Morgan, D.W., Bates, B. & Skinner, J.S. (1987), 'Physiological responses of triathletes to maximal swimming, cycling, and running.', *Med. Sci. Sports Exerc.* **19**, 51-55.

Machado, J.A.F. (1993), 'Robust Model Selection and M-estimation', *Econ-Ther.* **9**, 478-493.

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

Miller, A.J. (1990), *Subset Selection in Regression*, New York: Chapman and Hall.

Qian, G., & Künsch, H. (1996), 'On model selection in robust linear regression', *Res. rep. No. 80, Seminar für Statistik, Swiss Federal Institute of Technology*, Zürich (ETH). To appear in *J. Stat. Plan. & Infer.*.

Qian, G., & Künsch, H. (1998), 'Some notes on Rissanen's stochastic complexity.', *IEEE Trans. Inform. Theory.* **44**, 782-786.

Rao, C.R. & Wu, Y. (1989), 'A strongly consistent procedure for model selection in a regression problem', *Biometrika* **76**, 369-374.

Rissanen, J. (1986), 'Stochastic complexity and modeling', *Annals of Statistics*, **14**, 3, 1080–1100.

Rissanen, J. (1987), 'Stochastic complexity (with discussion)', *J. R. Statist. Soc., Ser. B*, **49**, 3, 223-265.

Rissanen J. (1989), *Stochastic Complexity in Statistical Inquiry*, World Scientific Publishing Co. Pte. Ltd., Singapore.

Rissanen, J. (1996), 'Fisher information and stochastic complexity', *IEEE Trans. Inform. Theory.* **42**, 40-47.

Ronchetti, E. (1985), 'Robust model selection in regression', *Stat. Prob. Lett.* **3**, 21-23.

Ronchetti, E. & Staudte, R.G. (1994), 'A robust version of Mallows's C_p', *J. Amer. Statist. Assoc.* **89**, 550-559.

Smith, A.F.M. & Roberts, G.O. (1993), 'Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods', *J. Roy. Statist. Soc. Ser. B* **55**, 3-23.

Tanner, M.A. (1996), *Tools for Statistical Inference*, 3rd Edition. Springer-Verlag, New York.

Venables, W.N. & Ripley, B.D. (1994), *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York.

Weisberg, S. (1985), *Applied Linear Regression* (2nd ed.), Wiley, New York.