Some Notes On Rissanen's Stochastic Complexity

by

Guoqi Qian [1] [2]
and
Hans R. Künsch

Research Report No. 79
November 1996

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

[1]Research supported by Swiss National Science Fund 2100-043210.95.
[2]Seminar für Statistik, ETH, and School of Statistics, La Trobe University, Melbourne 3083, Australia

Some Notes On Rissanen's Stochastic Complexity

Guoqi Qian [‡§]
and
Hans R. Künsch
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

November 1996

Abstract

A new version of stochastic complexity for a parametric statistical model is derived, based on a class of two-part codes. We show that choosing the quantization in the first step according to the Fisher information is optimal and we compare our approach to a recent result of Rissanen [10]. Application to robust regression model selection is presented.

Key words and phrases: stochastic complexity, model selection, robust regression.

1 Introduction

It is well known that the maximum of log likelihood alone cannot be used as a criterion for model selection since it would always select the full model. An additional term which penalizes the complexity of the model is needed. Various penalization terms have been proposed in the literature, based on widely different arguments. A particularly attractive concept is the stochastic complexity developed in the eighties by Rissanen and summarized in his book [9]. Stochastic complexity measures the goodness of fit of a model by its ability to compress the data. This is done by computing the length of a prefix code which has smallest expected length for the model under consideration. To a first approximation, this leads to the penalty term $-\frac{k}{2}\log n$ where k is the dimension of the model and n is the sample size. Recently in [10] Rissanen has derived a more precise approximation up to terms $o(1)$ as n tends to infinity. In this note we present a somewhat different coding procedure leading to a different approximation to stochastic complexity. It is the result of our attempt to understand a preliminary version of [10].

In order to describe the results, we have to introduce some notation. We denote the data by $x^n = (x_1, \cdots, x_n)$ and consider a statistical model $\mathcal{M} = \{f_n(x^n|\theta), \theta \in \Theta\}$ where $\Theta \subseteq \mathcal{R}^k$. Here each $f_n(x^n|\theta)$ is a probability density for x^n and the marginality condition holds:

$$\int f_{n+1}(x_1, \cdots, x_{n+1}|\theta)dx_{n+1} = f_n(x_1, \cdots, x_n|\theta).$$

If $\theta = \theta_0$, a fixed value, then the length of the optimal prefix code is known to be closely approximated by Shannon's complexity $-\log f_n(x^n|\theta_0)$. (In the sequel, the logarithm is base 2 by default.) Here we implicitly assume that x^n is observed to a certain precision, but its precise value does not matter.

If θ is unknown but belongs to Θ, then we can construct an optimal prefix code by a two-step encoding process, and use the resultant two-part code to describe x^n. Namely, we first encode the chosen member of $\mathcal{M}$, and then encode the data x^n with the optimal code for that member. The first step is equivalent to encoding the parameter space Θ. However, if Θ is uncountably infinite, there will be no code of finite length for describing Θ. One way to get over this difficulty is to truncate Θ to certain precision d by quantization. Then we encode the sequence of representatives of the quantized regions in Θ rather than the whole Θ. Denote Θ^d as the set of all representatives in Θ after the truncation and let L_d be the length of a prefix code for Θ^d. Defining

$$\hat{\theta}^d = \hat{\theta}^d(x^n) = \arg \min_{\theta^d \in \Theta^d} \{-\log f_n(x^n|\theta^d) + L_d(\theta^d)\}, \tag{1}$$

the optimal two-part code length for describing x^n relative to $\mathcal{M}$ is

$$-\log f_n(x^n|\hat{\theta}^d) + L_d(\hat{\theta}^d). \tag{2}$$

It is easy to see that (2) is indeed the length of a prefix code, i.e. Kraft's inequality is satisfied.

The finer the quantization is, the closer is $-\log f_n(x^n|\hat{\theta}^d)$ to the minimum of minus log likelihood, i.e. to $-\log f_n(x^n|\hat{\theta})$ where $\hat{\theta}$ is the MLE, but $L_d(\hat{\theta}^d)$ is larger. The opposite situation occurs if the quantization of Θ gets coarser and coarser. Thus an optimization problem for the quantization arises. It involves not only the value of d, but also the metric underlying the quantization. Typically the optimal quantization depends on x^n, but we require a choice which is independent of the data because otherwise we would need another code for the chosen quantization. We determine the quantization which is optimal if the data are generated according to some member of $\mathcal{M}$. But our expression remains a valid code length without such an assumption. It is given by

$$\begin{aligned} -\log f_n(x^n|\hat{\theta}) + \frac{k}{2}\log\frac{\rho_{n1}e}{4\log e} + \frac{1}{2}\log|I_n(\hat{\theta})| + \sum_{i=1}^{k}\log(|\hat{\theta}_i| + n^{-1/4}) \\ + \sum_{i=1}^{k} r^*(n^{1/4}|\hat{\theta}_i| + 1) + k(const + 1). \end{aligned} \tag{3}$$

Here θ_i is the i-th component of θ, $r^*(x) = \log(\log(x)) + \log(\log(\log(x)) + \ldots$ where the sum continues as long as the iterated logarithms are positive and $const \approx 2.87$. Furthermore

$$I_n(\theta) = E_\theta[-\frac{\partial^2 \log f_n(x^n|\theta)}{\partial\theta\partial\theta^t}] \tag{4}$$

is the expected Fisher information and ρ_{n1} is the maximal eigenvalue of $I_n(\hat{\theta})^{-\frac{1}{2}} J_n(x^n)$ $I_n(\hat{\theta})^{-\frac{1}{2}}$ where

$$J_n(x^n) = -\frac{\partial^2 \log f_n(x^n|\theta)}{\partial\theta\partial\theta^t}|_{\theta=\hat{\theta}} \tag{5}$$

is the observed Fisher information.

For simplicity we propose to retain in (3) only the leading terms. Then we obtain the following expression which we call approximate stochastic complexity of the data x^n with respect to the model $\mathcal{M}$

$$SC(x^n|\mathcal{M}) = -\log f_n(x^n|\hat{\theta}) + \frac{1}{2}\log|I_n(\hat{\theta})| + \sum_{i=1}^{k}\log(|\hat{\theta}_i| + n^{-\frac{1}{4}}). \tag{6}$$

In statistics, log is usually replaced by the natural logarithm ln. This will result in a change in the stochastic complexity by an additive term $-\frac{k}{2}\ln\ln 2$ and a multiplicative factor $\ln^{-1} 2$. Thus the change is minor.

For comparison, the stochastic complexity given in [10] is

$$SC(x^n|\mathcal{M}) = -\log f_n(x^n|\hat{\theta}) + \frac{k}{2}\log\frac{1}{2\pi} + \log\int_{\Theta(\hat{a})}\sqrt{|I_n(\theta)|}d\theta + \log^*(\hat{a}) + const. \tag{7}$$

Here $\Theta(1) \subset \Theta(2) \subset \ldots$ is an increasing sequence of bounded open subsets converging to Θ, $\hat{a}$ is the smallest integer a such that $\hat{\theta} \in \Theta(a)$ and $\log^*(x) = \log(x) + r^*(x)$. We will comment on the difference in section 3 below. Both formulae simplify in the case where Θ and $I(\theta)$ are bounded. But since our main interest is in regression problems where $\Theta = \mathcal{R}^k$, we concentrate here on the unbounded case.

2 Derivation

In order to make notation easier, we take $\Theta = \mathcal{R}^k$. We encode the parameter θ in two steps. First we use a coarse, but uniform discretization of Θ and a code length depending on $||\theta||$. In a second step we use a finer discretization with locally varying size and shape, but with uniform code length. The discretization procedure is partly used in [8]. For the first step we cover Θ by congruent cubes

$$C_\alpha \equiv C_\alpha(\Delta_n) = \{\theta \in \Theta \mid \Delta_n(|\alpha_i| - 1) \leq |\theta_i| \leq \Delta_n|\alpha_i|, sign(\alpha_i) = sign(\theta_i) \ (i = 1, \ldots k)\}$$

where $\alpha \in \{\pm 1, \pm 2, \ldots\}^k$, and the side width $\Delta_n > 0$ depending on the number of observations n satisfies $\lim_{n\to\infty}\Delta_n = 0$. Denoting by θ_α the center of C_α, we let $R(\theta_\alpha)$ be the maximal hyperrectangle which is contained in the ellipsoid

$$(\theta - \theta_\alpha)^t M_n(\theta_\alpha)(\theta - \theta_\alpha) \leq d. \tag{8}$$

Here $M_n(\theta)$ is a $k \times k$ positive definite matrix satisfying the following conditions:

(a). $M_n(\theta) > 0$.

(b). The minimum eigenvalue of $M_n(\theta)$ is of order $O(n)$ as $n \to \infty$.

(c). $|M_n(\theta_1)|^{-1}||M_n(\theta_1)| - |M_n(\theta_2)|| \leq c||\theta_1 - \theta_2||$ for any $\theta_1, \theta_2 \in \Theta$, where $c > 0$ being a finite constant.

(d). $\log|M_n(\theta)| = o(n)$.

Denoting the eigenvalues of $M_n(\theta_\alpha)$ by $\lambda_{n1,\alpha} \geq \lambda_{n2,\alpha} \cdots \geq \lambda_{nk,\alpha} > 0$, the length of the i-th side of $R(\theta_\alpha)$ is $2\sqrt{d/(k\lambda_{ni,\alpha})}$ and the volume of $R(\theta_\alpha)$ is

$$V(\theta_\alpha) = (4dk^{-1})^{\frac{k}{2}}|M_n(\theta_\alpha)|^{-\frac{1}{2}}. \tag{9}$$

Now starting from $R(\theta_\alpha)$ we cover C_α by a sequence of non-intersecting hyperrectangles which are translates of $R(\theta_\alpha)$. The resulting cover of C_α's is denoted by $\{R_{\alpha,\nu}(d), \nu = 1, \ldots N_\alpha\}$ and the centers by $\theta_{\alpha,\nu}(d)$. Thus encoding θ will be done by encoding $\Theta^d = \{\theta_{\alpha,\nu}(d)\}$, which is consequently equivalent to encoding ν and α. The sizes Δ_n and d as well as the matrices $M_n(\theta)$ determining the shape of the rectangles will be chosen later to minimize the resulting code length. Note that if $M_n(\theta)$ is the identity matrix times a constant, the two-step discretization is simply a uniform discretization of Θ.

For encoding ν given α we use an equal length code which has a length closely approximated by $\log(N_\alpha)$ (see [9], section 2.4). It is easy to get a lower and an upper bound for N_α as follows

$$\frac{\Delta_n^k}{V(\theta_\alpha)} \leq N_\alpha \leq \frac{(\Delta_n + 4\sqrt{d/(k\lambda_{nk,\alpha})})^k}{V(\theta_\alpha)}.$$

By (9) we thus obtain

$$\Delta_n^k \left(\frac{k}{4d}\right)^{k/2} |M_n(\theta_\alpha)|^{1/2} \leq N_\alpha \leq \Delta_n^k (1 + 4\sqrt{d/(k\lambda_{nk,\alpha})}\Delta_n^{-1})^k \left(\frac{k}{4d}\right)^{k/2} |M_n(\theta_\alpha)|^{1/2}. \quad (10)$$

For encoding α we need basically a code for the natural numbers. We use the so-called universal code ([1], [2], [6], [8]). It has the length $L(n) = \log^*(n) + const$ where $const \approx 2.87$ is determined by $\sum 2^{-L(n)} = 1$. Since we need an additional bit for the sign of α_i we obtain as the code length for α

$$\sum_{i=1}^{k} \log^*(|\alpha_i|) + k(const + 1).$$

From the definition of C_α, it follows that

$$\Delta_n^{-1}|\theta_{\alpha,\nu}(d)_i| \leq |\alpha_i| \leq \Delta_n^{-1}|\theta_{\alpha,\nu}(d)_i| + 1, \quad (11)$$

Taking (10) and (11) together we obtain the following upper bound of code length for encoding $\{\theta_{\alpha,\nu}(d)\}$:

$$\begin{aligned} L(\theta_{\alpha,\nu}(d)) \quad \leq \quad & \sum_{i=1}^{k} \log(|\theta_{\alpha,\nu}(d)_i| + \Delta_n) + \sum_{i=1}^{k} r^*(\Delta_n^{-1}|\theta_{\alpha,\nu}(d)_i| + 1) + \\ & \frac{k}{2}\log\frac{k}{4d} + \frac{1}{2}\log|M_n(\theta_\alpha)| + \frac{4\sqrt{dk}}{\Delta_n\sqrt{\lambda_{nk,\alpha}}} + k(const + 1). \end{aligned} \quad (12)$$

Provided that $\theta_{\alpha,\nu}(d)_i \neq 0$ the optimal order of Δ_n is by condition (b) $\Delta_n \geq O(n^{-1/4})$. Since it is not possible to be more precise, we will take $\Delta_n = n^{-1/4}$ by noting that a finer quantization loses less information. With this choice and with condition (c), we then obtain

$$\begin{aligned} L_d(\theta_{\alpha,\nu}(d)) \quad = \quad & \sum_{i=1}^{k} \log(|\theta_{\alpha,\nu}(d)_i| + n^{-1/4}) + \frac{k}{2}\log\frac{k}{4d} + \frac{1}{2}\log|M_n(\theta_{\alpha,\nu}(d))| + \\ & \sum_{i=1}^{k} r^*(n^{1/4}|\theta_{\alpha,\nu}(d)_i| + 1) + k(const + 1) + O(n^{-1/4}). \end{aligned} \quad (13)$$

From (2) and (13), the two-part code length for describing x^n relative to $\mathcal{M}$ is approximated by

$$\begin{aligned} L(x^n, \hat{\theta}^d) \quad \stackrel{def}{=} \quad & -\log f_n(x^n|\hat{\theta}^d) + \sum_{i=1}^{k} \log(|\hat{\theta}_i^d| + n^{-1/4}) + \frac{k}{2}\log\frac{k}{4d} + \\ & \frac{1}{2}\log|M_n(\hat{\theta}^d)| + \sum_{i=1}^{k} r^*(n^{1/4}|\hat{\theta}_i^d| + 1) + k(const + 1). \end{aligned} \quad (14)$$

Now we are going to determine the optimal d and M_n. By definition $\hat{\theta}^d$ is the value minimizing the expression (14). But we may as well take $\hat{\theta}^d$ to be the value closest to the maximum likelihood estimator $\hat{\theta}$ since this can only increase the code length. Hence we will do so. Actually in Lemma 1 at the end of this section, we will show that this changes $\hat{\theta}^d$ by $O(n^{-1/2})$ under suitable conditions. Hence the increase of the code length is of the order $O(1)$. By a Taylor expansion we obtain

$$-\log f_n(x^n|\hat{\theta}^d) = -\log f_n(x^n|\hat{\theta}) + \frac{1}{2}(\hat{\theta}^d - \hat{\theta})^t J_n(x^n)(\hat{\theta}^d - \hat{\theta})(1 + o(1))$$

where $J_n(x^n)$ is the observed Fisher-information defined in (5). By a standard result from linear algebra we have for any vector $z \in R^k$

$$z^t J_n z \leq \rho_{n1} z^t M_n(\hat{\theta}) z$$

where $\rho_{n1} \geq \ldots \geq \rho_{nk}$ are the eigenvalues of $M_n(\hat{\theta})^{-1/2} J_n M_n(\hat{\theta})^{-1/2}$. Hence we obtain the following upper bound for (14)

$$\begin{aligned} -\log f_n(x^n|\hat{\theta}) + \frac{1}{2}\rho_{n1} d + \sum_{i=1}^{k} \log(|\hat{\theta}_i| + n^{-1/4}) + \frac{k}{2}\log\frac{k}{4d} + \frac{1}{2}\log|J_n(x^n)| - \\ \frac{1}{2}\log(\prod_{i=1}^{k}\rho_{ni}) + \sum_{i=1}^{k} r^*(n^{1/4}|\hat{\theta}_i| + 1) + k(const + 1) + O(n^{-1/4}). \end{aligned} \tag{15}$$

It is easy to see that the optimal d and M_n minimizing (15) are

$$M_n(\hat{\theta}) = J_n, \quad d = k \log e.$$

But this creates problems because typically there exist x^n and z^n such that $\hat{\theta}(x^n) = \hat{\theta}(z^n)$, but $J_n(x^n) \neq J_n(z^n)$. For a code length, M_n must not depend on the data (otherwise we have to encode also M_n). But if the data are distributed according to $f_n(x^n|\theta)dx^n$ for some θ, then the observed Fisher information J_n is approximately equal to the expected Fisher information I_n defined in (4). We therefore propose to always use $M_n(\theta) = I_n(\theta)$ regardless whether the data come from the model or not. The possibility that the data are not generated by one of our model distributions and thus I_n may differ substantially from J_n is taken into account by computing ρ_{n1}, the maximal eigenvalue of $I_n(\hat{\theta})^{-1/2} J_n(x^n) I_n(\hat{\theta})^{-1/2}$.

Taking these results together we find the approximate code length to be given by (3). Note that we require only that the Fisher information I_n exists and satisfies conditions (a) - (d) from the beginning of this section. The difference between (3) and the approximation (6) is $O(k \log\log n) + O(k \log \rho_{n1})$. Note that typically $n^{-1} I_n(\theta)$ and $n^{-1} J_n(x^n)$ both have a limit and thus $\rho_{n1} = O(1)$.

We conclude this section by listing the following lemma mentioned before:

Lemma 1 *In addition to conditions (a) to (d), suppose that $f_n(x^n|\theta)$ is three times continuously differentiable with respect to θ and $J_n(x^n) > 0$ a.s. satisfying $\lambda_k(J_n) = O(n)$ a.s. where $\lambda_k(\cdot)$ denotes the minimum eigenvalue. Then we have $|\hat{\theta} - \hat{\theta}^d| = O(n^{-1/2})$ a.s. where $\hat{\theta}^d$ is defined by (1).*

Proof: Denote $T_n(\theta) = L(x^n, \theta) + \log f_n(x^n|\theta)$. It is easy to see from condition (d) and (14) that $T_n(\theta) = o(n)$. Now make a Taylor expansion for $-\log f_n(x^n|\hat{\theta}^d)$ around $\hat{\theta}$,

$$-\log f_n(x^n|\hat{\theta}^d) = -\log f_n(x^n|\hat{\theta}) + \frac{1}{2}(\hat{\theta}^d - \hat{\theta})^t J_n(x^n)(\hat{\theta}^d - \hat{\theta})(1 + o(1)) \quad \text{a.s.} \tag{16}$$

since $-\log f_n(x^n|\theta)$ is three times continuously differentiable with respect to θ. It follows that

$$L(x^n, \hat{\theta}^d) = L(x^n, \hat{\theta}) + T_n(\hat{\theta}^d) - T_n(\hat{\theta}) + \frac{1}{2}(\hat{\theta}^d - \hat{\theta})^t J_n(x^n)(\hat{\theta}^d - \hat{\theta})(1 + o(1)) \quad \text{a.s..}$$

Thus

$$(\hat{\theta}^d - \hat{\theta})^t J_n(x^n)(\hat{\theta}^d - \hat{\theta}) \leq 2(T_n(\hat{\theta}) - T_n(\hat{\theta}^d)) \quad \text{a.s.} \tag{17}$$

if n is sufficiently large, since $L(x^n, \hat{\theta}^d) \leq L(x^n, \hat{\theta})$. Using the property $T_n(\theta) = o(n)$ and the condition for $J_n(x^n)$, we obtain $|\hat{\theta}^d - \hat{\theta}| = o(1)$ a.s. as $n \to \infty$.

From this result and condition (c) it readily follows that $0 \leq T_n(\hat{\theta}) - T_n(\hat{\theta}^d) \leq O(1)$ a.s.. Using this inequality and (17) again, it follows that $|\hat{\theta}^d - \hat{\theta}| = O(n^{-1/2})$ a.s.. □

3 Discussion

Our approach differs from the one in [10] in two aspects. The first is that we do not remove an inherent redundancy of the two-part code (2). After encoding $\hat{\theta}^d$ it would be sufficient to give a code only for those x^n which have the same solution to the minimization problem (1). In the notation of section 2 we could thus use the code length

$$-\log f(x^n|\theta_{\alpha,\nu}(d)) + \log P[\hat{\theta} \in R_{\alpha,\nu}(d)|\theta_{\alpha,\nu}(d)] + L(\theta_{\alpha,\nu}(d)).$$

But for $M_n = I_n$, typically $M_n = O(n)$ and the gain is only of $O(1)$. Moreover its computation is complicated involving multivariate normal probabilities with a general covariance matrix. So we omit it. Rissanen(1996) uses $M_n = nr_n^{-2}Id$ where I is the identity matrix and $r_n = o(1)$. With such a choice

$$\log P[\hat{\theta} \in R_{\alpha,\nu}(d)|\theta_{\alpha,\nu}(d)] \to -\infty$$

which is a paradox since even with infinite precision for $\hat{\theta}$ we still need to encode x^n. Note that the precise value of $\hat{\theta}$ restricts x^n typically to a $(n-k)$-dimensional manifold. The resolution of the paradox lies in the precision δ for the data x^n. Knowing $\hat{\theta}$, we need typically a code length of $O(-(n-k)\log\delta)$ to encode x^n whereas without this information we need a length of $O(-n\log\delta)$. So removing the redundancy when knowing $\hat{\theta}$ with infinite precision reduces the length by $O(-k\log\delta)$. This is arbitrarily large only for δ going to zero. The same conclusion is obtained by the following argument. The expression

$$-n\log\delta - \log f(x^n|\theta_{\alpha,\nu}(d)) + \log P[\hat{\theta} \in R_{\alpha,\nu}(d)|\theta_{\alpha,\nu}(d)]$$

gives the approximate code length for encoding x^n to precision δ subject to knowing $\hat{\theta} \in R_{\alpha,\nu}(d)$ only if the width of that set is larger than δ. As the case of i.i.d. normal random variables shows, that width shrinks to zero if M_n grows faster than n. So we believe that there are good reasons not to let M_n grow faster than n.

The second difference between our approach and the one of [10] is the way unbounded parameters are handled. It seems inevitable to introduce a two-stage procedure where one uses the universal prior for integers in the first, coarse stage. In [10] large sets are chosen in this first stage whereas in our approach the size Δ_n of the first stage partition tends to zero.

In addition to the minimum of minus log likelihood, our criterion (6) contains two penalty terms. Let us briefly discuss their role. Note that if all the components of θ are quite away from zero, so will be their estimates for n sufficiently large since the maximum likelihood estimator is consistent. Thus in this case the term $\sum \log(|\hat{\theta}_i|+n^{-1/4})$ is negligible compared to the other term $\frac{1}{2}\log|I_n(\hat{\theta})|$. On the other hand, however, if any $\theta_i = 0$, it follows that $\hat{\theta}_i = O_p(n^{-1/2})$ assuming that $\hat{\theta}_i$ satisfies the central limit theorem. Then $\log(|\hat{\theta}_i| + n^{-1/4}) \sim -\frac{1}{4}\log n$ is comparable to $\frac{1}{2}\log|I_n(\hat{\theta})| = O(\frac{k}{2}\log n)$ and it reduces the stochastic complexity. Therefore our criterion favors somewhat models which contain estimated parameters close to zero, but not so much that the procedure becomes inconsistent.

An unsatisfactory feature of both Rissanen's and our criterion is that they are not automatically invariant under reparameterization. In Rissanen's case the problem comes from the choice of the bounded open sets $\Theta(a)$ where it is not clear how to choose them. Under a reparameterization one has to transform these sets accordingly. For our criterion the most important point is where one puts the origin in the parameter space. One should choose the parameterization such that setting $\theta_i = 0$ corresponds to a simpler model for each i. Usually this can be done in some canonical way. In order to see how a reparameterization affects our criterion once the origin is fixed, choose $\theta_i = h_i(\xi_1, \cdots, \xi_k), \quad i = 1, \cdots, k$, such that $\xi_i = 0$ if and only if $\theta_i = 0$. Then the change in our criterion (6) is

$$\log\left|\left(\frac{\partial\theta_s}{\partial\xi_t}\right)\right|_{\hat{\theta}} + \sum_{j=1}^{k}\log\frac{|\hat{\xi}_j| + n^{-1/4}}{|\hat{\theta}_i| + n^{-1/4}}.$$

This change is bounded by $O(k)$ under general conditions. Therefore, the stochastic complexity (6) as a model selection criterion is not affected for large sample situation and mildly affected otherwise under reparameterization.

A final problem is the term $n^{-1/4}$ in our criterion (6). From section 2 we know that it is obtained by an approximation. But when changing the scale of θ_i we should also scale this $n^{-1/4}$ term with it. This can be done for example by replacing $|\theta_i| + n^{-1/4}$ with $|\theta_i| + \varepsilon_i n^{-1/4}$ where $\varepsilon_i = (n^{-1}I_n(0)_{ii})^{-1/2}$ and $I_n(0)_{ii}$ is the i-th diagonal element of $I_n(0)$. We illustrate this by example 2 in section 4.

4 Examples

Example 1: *Exponential Variables.* Suppose $x_1, \cdots, x_n$ are the outcomes of n i.i.d. exponential variables with density $f(x|\theta) = \theta\exp(-\theta x)$. In this case a canonical parameterization is given by $\xi = \ln(\theta)$ and $\xi = 0$ is a reasonable choice of the origin. Simple calculations lead to $\hat{\xi} = -\ln(\overline{x})$ and $J_n(x^n) \equiv I_n(\xi) \equiv n\log e$. Therefore the precise expression (3) is in this case given by

$$n\log e(1 + \ln\overline{x}) + \frac{1}{2}\log e + \frac{1}{2}\log n + \log(|\ln\overline{x}| + n^{-1/4}) + r^*(n^{1/4}|\ln\overline{x}| + 1) + const.$$

The simplified expression (6) using the natural logarithm ln becomes

$$SC(x^n|\mathcal{M}) = n(1 + \ln\overline{x}) + \frac{1}{2}\ln n + \ln(|\ln\overline{x}| + n^{-1/4}).$$

Example 2: *Robust Regression.* In regression, the aim is to model the linear dependence of responses y_i on a p-dimensional explanatory variable x_i

$$y_i = x_i^t\beta + r_i. \tag{18}$$

Here β is a p-dimensional unknown parameter and r_i are the errors. For model selection, we consider submodels obtained by setting certain components of β equal to zero. The classical method takes the errors to be i.i.d. $\mathcal{N}(0,\sigma^2)$-distributed which gives least squares as the MLE. But least squares are known to be non-robust, that is changes in one or a few observations can lead to entirely different estimates $\hat{\beta}$. A similar instability occurs in the model selection based on least squares. In the language of stochastic complexity this means that the optimal code for Gaussian errors can be very long if one or a few observations are not typical for such a model. One could try to treat the distribution of the errors as an infinite dimensional nuisance parameter and to estimate it using again ideas from stochastic complexity. However, such infinite dimensional nuisance parameters are delicate both in theory and practice, see e.g. [3] and [11]. The approach of robust statistics is different: It uses a single distribution which should give a reasonably short code length for all data which are typical for an arbitrary distribution in an infinite dimensional neighborhood of the normal. The most famous distribution to achieve this is the so-called least favorable distribution of Huber, see [5]. We refer the reader to that paper for some theory concerning its optimality.

An additional problem occurs when an explanatory variable x_i is very different from the rest. Such an observation is highly influential even with the least favorable distribution, cf. [4], Section 6.2. Our way to deal with this problem is to employ in addition a varying scale depending on how far out x_i is. Thus in our model the r_i's are independent and have the density

$$f(r_i) = (1-\lambda)(\sqrt{2\pi}\sigma)^{-1} w_i \exp\{-\rho_c(\frac{w_i r_i}{\sigma})\} = \frac{w_i}{\sigma} f_0(\frac{w_i r_i}{\sigma}), \tag{19}$$

where $f_0(r) = (1-\lambda)(\sqrt{2\pi})^{-1}\exp\{-\rho_c(r)\}$ with $0 < \lambda < 1$, σ is a scale parameter, and $w_i = w(x_i) \in (0,1]$ a weight function. The function $\rho_c(\cdot)$ is the Huber function defined as

$$\rho_c(t) = \begin{cases} \frac{1}{2}t^2, & |t| < c \\ c|t| - \frac{1}{2}c^2, & |t| \geq c. \end{cases}$$

The constants λ and c are related by $(1-\lambda)^{-1} = 2\Phi(c) - 1 + 2\phi(c)/c$ where $\Phi(\cdot)$ is the distribution function of $\mathcal{N}(0,1)$ and $\phi(\cdot)$ is its density. It easily follows that

$$E(r_i|x_i) = 0, \quad \mathrm{Var}(r_i|x_i) = \frac{\sigma^2}{w_i^2}(1-\lambda)\{2\Phi(c) - 1 + 4(\frac{1}{c} + \frac{1}{c^3})\phi(c)\}.$$

Thus observations with a small weight have a large (conditional) variance in our model and thus a reduced influence on the MLE. Apparently the weights have to be small for outlying x_i's. The choice of $w(x)$ is discussed further in [7]. There it is also shown that the influence function of the MLE is indeed bounded if $w(x)||x||$ is bounded.

If our ultimate aim is to select an optimal predictor in the regression model, σ can be treated as a nuisance parameter and we do not need to spend any codeword to describe it in the stochastic complexity that we will derive. In practice σ can be replaced by a robust estimate using, for instance, Huber's proposal 2 or Hampel's median absolute deviation with a little modification for the full model (see [7]).

From (19), the (conditional) log-likelihood given the explanatory variables $X_n = (x_1, \cdots, x_n)^t$ has the form

$$\ell(y^n|X_n,\beta) = n\ln(1-\lambda) - \frac{n}{2}\ln 2\pi - n\ln\sigma + \sum_{i=1}^n \ln w_i - \sum_{i=1}^n \rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\beta)\}. \tag{20}$$

From this we can easily compute the observed and the expected Fisher information matrix:

$$J_n(y^n|X_n) = -\frac{\partial^2\ell}{\partial\beta\partial\beta^t} = \frac{1}{\sigma^2}\sum_{i=1}^n \rho_c''\{\frac{w_i}{\sigma}(y_i - x_i^t\beta)\} w_i^2 x_i x_i^t$$

where $\rho_c^{''} = 0$ for $|t| \geq c$ and 1 for $|t| < c$. Thus we obtain

$$I_n(\beta|X_n) = -E(\frac{\partial^2 \ell}{\partial\beta\partial\beta^t}) = \frac{1}{\sigma^2}(E_{f_0}\rho_c^{''})X_n^t W_n^2 X_n = \frac{1}{\sigma^2}\frac{2\Phi(c)-1}{2\Phi(c)-1+2c^{-1}\phi(c)}X_n^t W_n^2 X_n, \tag{21}$$

where $W_n = diag(w_1, \cdots, w_n)$. Note that the maximal eigenvalue ρ_{n1} is always bounded by a finite number. Thus we will omit it in the following.

Following (6) and (18) to (21), the stochastic complexity of Y_n relative to the regression model (18) and the (conditional) least favorable distributions for r_i's can be well approximated by

$$SC(y^n|X_n,\sigma) = -\ell(y^n|X_n,\hat{\beta}) + \frac{1}{2}\ln|I_n(\hat{\beta}|X_n)| + \sum_{i=1}^{p}\ln(|\hat{\beta}_i| + n^{-1/4})$$

$$= \sum_{i=1}^{n}\rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\hat{\beta})\} + \frac{p}{2}\ln E_{f_0}\rho_c^{''} + \frac{1}{2}\ln|X_n^t W_n^2 X_n| + \ln\prod_{i=1}^{p}\frac{|\hat{\beta}_i| + n^{-1/4}}{\sigma} \tag{22}$$

plus terms irrelevant to model selection provided that w_i's are determined from the independent variables in the full model (see [7]). By the principle of minimum description length, a criterion for selecting the explanatory variables in robust linear regression is obtained by computing the approximation (22) for all submodels obtained by omitting columns of X_n and choosing the one which minimizes this criterion.

Let us give some interpretations for (22). The first term in (22) is the sum of the robustified prediction errors which shows the goodness of the robust fit to the observations. The others represent model complexity. The second term gives a cost for using a robust method. The third one gives the weighted magnitude of the explanatory variables and the last one the (generalized) signal-to-noise ratio. This greatly extends previous formulation of the model complexity which depend only on the dimension of the parameter, e.g. in AIC, BIC and Mallows' C_p.

Finally we propose a modification to make (22) invariant. We assume that $x_{i1} \equiv 1$, i.e. the regression contains an intercept. In this situation, it is desirable that the criterion is invariant under both location and scale transformations of y and of the last $p-1$ components of x (assumed to be linearly independent). It is easily seen that this is achieved if we modify (22) as follows:

$$\begin{aligned} SC^{'}(y^n|X_n,\sigma) &= \sum_{i=1}^{n}\rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\hat{\beta})\} + \frac{p}{2}\ln E_{f_0}\rho_c^{''} \\ &+ \frac{1}{2}\ln|X_n^t W_n^2 X_n| + \ln\prod_{j=2}^{p}\left(\frac{|\hat{\beta}_j|}{\sigma} + s_{x(j)}^{-1}n^{-1/4}\right). \end{aligned} \tag{23}$$

where

$$s_{x(j)}^2 = (\sum_{i=1}^{n} w_i^2)^{-1}\sum_{i=1}^{n} w_i^2(x_{ij} - \bar{x}_{\cdot j})^2, \quad \bar{x}_{\cdot j} = (\sum_{i=1}^{n} w_i^2)^{-1}\sum_{i=1}^{n} w_i^2 x_{ij}$$

For more details about this criterion we refer to [7].

References

[1] J.L. Bentley and A.C. Yao, "An almost optimal algorithm for unbounded searching," *Inform. Processing Letters* **5**, 82-87, 1976.

[2] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory* **21**, 194-203, 1975.

[3] P. Hall and E.J. Hannan, "On stochastic complexity and nonparametric density estimation," *Biometrika* **75**, 705-714, 1988.

[4] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel. *Robust Statistics, The Approach Based On Influence Functions.* Wiley, New York, 1986.

[5] P.J. Huber, "Robust estimation of a location parameter," *Ann. Math. Stat.* **35**, 73-101, 1964.

[6] S.K. Leung-Yan-Cheong and T. Cover, "Some equivalences between Shannon entropy and Kolmogorov complexity," *IEEE Trans. Inform. Theory* **24**, 331-338, 1978.

[7] G. Qian and H.R. Künsch, "On Model Selection in Robust Linear Regression," preprint, 1996.

[8] J. Rissanen, "A universal prior for integers and estimation by minimum description length" *Ann. Statist.* **11**, 416-431, 1983.

[9] J. Rissanen, *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Co. Pte. Ltd., Singapore, 1989.

[10] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory* **42**, 40-47, 1996.

[11] B. Yu and T. Speed, "Data compression and histograms," *Probab. Theory Related Fields* **92**, 195-229, 1992.