Efficient Computation of Stochastic Complexity

Petri Kontkanen, Wray Buntine, Petri Myllymäki, Jorma Rissanen, Henry Tirri
Complex Systems Computation Group (CoSCo), Helsinki Institute for Information Technology (HIIT)
University of Helsinki & Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland.
{Firstname}.{Lastname}@hiit.fi

Abstract

Stochastic complexity of a data set is defined as the shortest possible code length for the data obtainable by using some fixed set of models. This measure is of great theoretical and practical importance as a tool for tasks such as model selection or data clustering. Unfortunately, computing the modern version of stochastic complexity, defined as the Normalized Maximum Likelihood (NML) criterion, requires computing a sum with an exponential number of terms. Therefore, in order to be able to apply the stochastic complexity measure in practice, in most cases it has to be approximated. In this paper, we show that for some interesting and important cases with multinomial data sets, the exponentiality can be removed without loss of accuracy. We also introduce a new computationally efficient approximation scheme based on analytic combinatorics and assess its accuracy, together with earlier approximations, by comparing them to the exact form. The results suggest that due to its accuracy and efficiency, the new sharper approximation will be useful for a wide class of problems with discrete data.

1 INTRODUCTION

From the information-theoretic point of view, the most plausible explanation for a phenomenon is the one which can be used for constructing the most effective coding of the observable realizations of the phenomenon. This type of *minimum encoding* explanations can be applied in statistical learning for building realistic domain models, given some sample data. Intuitively speaking, in principle this approach can be argued to produce the best possible model of the problem domain, since in order to be able to produce the most efficient coding of data, one must capture all the regularities present in the domain. Consequently, the minimum encoding approach can be used for constructing a solid theoretical framework for statistical modeling. Similarly, the minimum encoding approach can be used for producing accurate predictions of future events.

The most well-founded theoretical formalization of the intuitively appealing minimum encoding approach is the *Minimum Description Length (MDL)* principle developed by Rissanen (Rissanen, 1978, 1987, 1996). The MDL principle has gone through several evolutionary steps during the last two decades. For example, the early realization of the MDL principle, the two-part code MDL (Rissanen, 1978), takes the same form as the Bayesian BIC criterion (Schwarz, 1978), which has led some people to incorrectly believe that MDL and BIC are equivalent. The latest instantiation of MDL discussed here is *not* directly related to BIC, but to a more evolved formalization described in (Rissanen, 1996). For discussions on the theoretical advantages of this approach, see e.g. (Rissanen, 1996; Barron, Rissanen, & Yu, 1998; Grünwald, 1998; Rissanen, 1999; Xie & Barron, 2000; Rissanen, 2001) and the references therein.

The most important notion of MDL is the *Stochastic Complexity (SC)*, which is defined as the shortest description length of a given data relative to a model class $\mathcal{M}$. Unlike some other approaches, like for example Bayesian methods, the MDL principle does not assume that the model class chosen is correct. It even says that there is no such thing as a true model or model class, which in Bayesian methods is sometimes acknowledged in practice. Furthermore, SC is an objective criterion in the sense that it is not dependent on any prior distribution, it only uses the data at hand[1]. This means that the objectives of the MDL approach are very similar to those behind Bayesian methods with so-called reference priors (Bernardo, 1997), but note, however, that Bernardo himself expresses doubt that a reasonably general notion of "non-informative" pri-

[1]Unlike Bayesian methods, with SC the possible subjective prior information is not used as an explicit part of the theoretical framework, but it is expected to be used implicitly in the selection of the parametric model class discussed in the next section.

ors exists in Bayesian statistics in the multivariate framework (Bernardo, 1997).

It has been shown (see (Clarke & Barron, 1990; Grünwald, 1998)) that the stochastic complexity criterion is asymptotically equivalent to the asymptote of the Bayesian marginal likelihood method with the Jeffreys prior under certain conditions, when the Jeffreys prior also becomes equivalent to the so-called reference priors (Bernardo & Smith, 1994). Nevertheless, with discrete data this equivalence does not hold near the boundary of the parameter space in many models (Chickering & Heckerman, 1997; Xie & Barron, 2000), and in applications such as document or natural language modelling some parameters are expected to lie at the boundary. The implicit use of the Laplace approximation in the Bayesian derivations severely strains the approximation or completely anulls it on the boundaries, as discussed in (Bernardo & Smith, 1994; Bleistein & Handelsman, 1975). Consequently, it can be said that the stochastic complexity approach aims to achieve the goal of objectivity in a way not demonstrated in the Bayesian approach due to technical difficulties.

All this makes the MDL principle theoretically very appealing. However, the applications of the modern, so called Normalized Maximum Likelihood (NML) version of MDL, at least with multinomial data, have been quite rare. This is due to the fact that the definition of SC involves a sum (or integral) over all the possible data matrices of certain length, which are obviously exponential in number. Some applications have been presented for discrete regression (Tabus, Rissanen, & Astola, 2002), linear regression (Barron et al., 1998; Dom, 1996), density estimation (Barron et al., 1998) and segmentation of binary strings (Dom, 1995). In this paper, we will present methods for removing the exponentiality of SC in several important cases involving multinomial (discrete) data. Even these methods are, however, in some cases computationally demanding. Therefore we also present three computationally efficient approximations to SC and instantiate them for the cases mentioned. The approach is similar to our previous work in (Kontkanen, Myllymäki, Silander, & Tirri, 1999), but it was based on an earlier definition of MDL, not on the modern version adopted here. The ability to compute the exact SC gives us a unique opportunity to see how accurate the approximations are. This is important as we firmly believe that the results extend to more complex cases where exact SC is not available.

In Section 2 we first review the MDL principle and discuss how to compute it for a single multinomial variable and a certain multi-dimensional model class. The techniques used in this section are completely new. Section 3 presents the three SC approximations for multinomial data. In Section 4 we study the accuracy of these approximations by comparing them to the exact stochastic complexity. Finally, Section 5 gives the concluding remarks and presents some ideas for future work.

2 STOCHASTIC COMPLEXITY FOR MULTINOMIAL DATA

2.1 INTRODUCTION TO MDL

Let us consider a data set (or matrix) $\mathbf{x}^N = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of N outcomes (vectors), where each outcome $\mathbf{x}_j$ is an element of the set $\mathcal{X}$. The set $\mathcal{X}$ consists of all the vectors of the form $(a_1, \ldots, a_m)$, where each variable (or attribute) a_i takes on values $v \in \{1, \ldots, n_i\}$. Furthermore, we assume that our data is multinomially distributed.

We now consider the case with a parametric family of *probabilistic candidate models* (or codes) $\mathcal{M} = \{f(\mathbf{x}|\theta) \mid \theta \in \Gamma\}$, where Γ is an open bounded region of $\mathbf{R}^k$ and k is a positive integer. The basic principle behind *Minimum Description Length (MDL)* modeling is to find a code that minimizes the code length over all data sequences which can be well modeled by $\mathcal{M}$. Here a data sequence being "well-modeled by $\mathcal{M}$" means that there is a model θ in $\mathcal{M}$ which gives a good fit to the data. In other words, if we let $\hat{\theta}(\mathbf{x}^N)$ denote the maximum likelihood estimator (MLE) of the data $\mathbf{x}^N$, then $\mathbf{x}^N$ is well modeled by $\mathcal{M}$ means that $f(\mathbf{x}^N|\hat{\theta}(\mathbf{x}^N))$ is high. The *stochastic complexity* of a data sequence $\mathbf{x}^N$, relative to a family of models $\mathcal{M}$, is the code length of $\mathbf{x}^N$ when it is encoded using the most efficient code obtainable with the help of the family $\mathcal{M}$.

In the above, stochastic complexity was defined only in an implicit manner — as discussed in (Grünwald, Kontkanen, Myllymäki, Silander, & Tirri, 1998), there exist several alternative ways for defining the stochastic complexity measure and the MDL principle explicitly. In (Rissanen, 1996) Rissanen shows how the two-part code MDL presented in (Rissanen, 1978) can be refined to a much more efficient coding scheme. This scheme is based on a notion of *normalized maximum likelihood (NML)*, proposed for finite alphabets in (Shtarkov, 1987). The definition of NML is

$$P_{NML}(\mathbf{x}^N \mid \mathcal{M}) = \frac{P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N), \mathcal{M})}{\sum_{\mathbf{y}^N} P(\mathbf{y}^N \mid \hat{\theta}(\mathbf{y}^N), \mathcal{M})}, \quad (1)$$

where the sum goes over all the possible data matrices of length N. For discussions on the theoretical motivations behind this criterion, see e.g. (Rissanen, 1996; Merhav & Feder, 1998; Barron et al., 1998; Grünwald, 1998; Rissanen, 1999; Xie & Barron, 2000; Rissanen, 2001).

Definition (1) is intuitively very appealing: every data matrix is coded using its own maximum likelihood (i.e. best fit) model, and then a penalty for the complexity of the model class $\mathcal{M}$ is added to normalize the distribution. This penalty, i.e., the denominator of (1), is called the *regret*. Note that usually the regret is defined as a logarithm of the

denominator. In this paper, however, we mostly use the language of probability theory rather than information theory and thus the definition without the logarithm is more natural.

2.2 COMPUTING THE NML: ONE-DIMENSIONAL CASE

We now turn to the question of how to compute the NML criterion (1), given a data matrix $\mathbf{x}^N$ and a model class $\mathcal{M}_1$. Let us first consider a case with only one multinomial variable with K values. The maximum likelihood term is easy and efficient to compute:

$$P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N), \mathcal{M}_1) = \prod_{j=1}^{N} P(\mathbf{x}_j \mid \hat{\theta}(\mathbf{x}^N)) = \prod_{v=1}^{K} \hat{\theta}_v^{h_v} = \prod_{v=1}^{K} \left(\frac{h_v}{N}\right)^{h_v}, \quad (2)$$

where $\hat{\theta}_v$ is the probability of value v, and $(h_1, \ldots, h_K)$ are the *sufficient statistics* of $\mathbf{x}^N$, which in the case of multinomial data are simply the frequencies of the values $\{1, \ldots, K\}$ in $\mathbf{x}^N$.

At first sight it may seem that the time complexity of computing the regret, i.e., the denominator in (1), grows exponentially with the size of the data, since the summing goes over K^N terms. However, it turns out that for reasonable small values of K it is possible to compute (1) efficiently. Since the maximum likelihood (2) only depends on the sufficient statistics h_v, the regret can be written as

$$R^1_{K,N} \stackrel{\text{def.}}{=} \sum_{\mathbf{x}^N} P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N), \mathcal{M}_1) = \sum_{h_1+\cdots+h_K=N} \frac{N!}{h_1! \cdots h_K!} \prod_{v=1}^{K} \left(\frac{h_v}{N}\right)^{h_v}, \quad (3)$$

where in the last formula the summing goes over all the *compositions* of N into K parts, i.e., over all the possible ways to choose non-negative integers $h_1, \ldots, h_K$ so that they sum up to N. We use the notation $R^1_{K,N}$ to refer to this subsequently, i.e., the regret for one multinomial variable with K values and N data vectors. The time complexity of (3) is $\mathcal{O}(N^{K-1})$, which is easy to see. For example, take case $K = 3$. The regret can be computed in $\mathcal{O}(N^2)$ time:

$$R^1_{3,N} = \sum_{h_1=0}^{N} \sum_{h_2=0}^{N-h_1} \frac{N!}{h_1! h_2! (N-h_1-h_2)!} \cdot \left(\frac{h_1}{N}\right)^{h_1} \left(\frac{h_2}{N}\right)^{h_2} \left(\frac{N-h_1-h_2}{N}\right)^{N-h_1-h_2}. \quad (4)$$

2.3 COMPUTING THE NML : THE RECURSIVE FORMULA

It turns out that the exact regret for a single multinomial variable can also be computed with a computationally very efficient combinatoric recursive formula. Consider $R^1_{K,N}$ as before. Using standard combinatorics we get the following recursion:

$$R^1_{K,N} = \sum_{h_1+h_2=N} \frac{N!}{h_1! h_2!} \left(\frac{h_1}{N}\right)^{h_1} \left(\frac{h_2}{N}\right)^{h_2} \cdot R^1_{k1,h1} R^1_{k2,h2}, \quad (5)$$

where $k_1 + k_2 = K$.

This formula allows us to compute the exact NML very efficiently by applying a common doubling trick from combinatorics. Firstly, one computes the tables of $R^1_{2^m,n}$ for $m = 1, \ldots, \lfloor \log K \rfloor$ and $n = 1, \ldots, N$. Secondly, $R^1_{K,N}$ can be built up from these tables. For example, take the case $R^1_{26,N}$. First calculate $R^1_{K,n}$ for $K \in \{2, 4, 8, 16\}$ and $n = 1, \ldots, N$. Then apply (5) to calculate the tables of $R^1_{10,n}$ from $R^1_{2,n}$ and $R^1_{8,n}$. Finally, $R^1_{26,N}$ can be computed from the tables of $R^1_{16,n}$ and $R^1_{10,n}$. It is now easy to see that the time complexity of computing (5) is $\mathcal{O}(N^2 \log K)$.

2.4 COMPUTING THE NML : MULTI-DIMENSIONAL CASE

The one-dimensional case discussed in the previous sections is not adequate for many real-world situations, where data is typically multi-dimensional. Let us assume that we have m variables. The number of possible data vectors is $\prod_{i=1}^{m} n_i$. It is clear that even the methods presented in the previous sections do not make the NML computation efficient in the multi-dimensional case. We are forced to make some independence assumptions. In this article, we assume the existence of a special variable c (which can be chosen to be one of the variables in our data matrix or it can be latent), and that given the value of c, the variables $(a_1, \ldots, a_m)$ are independent. That is, denoting the model class resulting from this assumption by $\mathcal{M}_T$,

$$P(c, a_1, \ldots, a_m \mid \theta, \mathcal{M}_T) = P(c \mid \theta, \mathcal{M}_T) \prod_{i=1}^{m} P(a_i \mid c, \theta, \mathcal{M}_T). \quad (6)$$

Although simple, this model class has been very successful in practice in mixture modeling (Kontkanen, Myllymäki, & Tirri, 1996), cluster analysis, case-based reasoning (Kontkanen, Myllymäki, Silander, & Tirri, 1998), Naive Bayes classification (Grünwald et al., 1998; Kontkanen, Myllymäki, Silander, Tirri, & Grünwald, 2000) and data visualization (Kontkanen, Lahtinen, Myllymäki, Silander, & Tirri, 2000).

We now show how to compute NML for $\mathcal{M}_T$. Assuming c has K values and using (3), Equation (1) becomes

$$P_{NML}(\mathbf{x}^N|\mathcal{M}_T) = \frac{\prod_{k=1}^{K}\left(\frac{h_k}{N}\right)^{h_k}\prod_{i=1}^{m}\prod_{k=1}^{K}\prod_{v=1}^{n_i}\left(\frac{f_{ikv}}{h_k}\right)^{f_{ikv}}}{R^m_{\mathcal{M}_T}}, \quad (7)$$

where h_k is the number of times c has value k in $\mathbf{x}^N$, f_{ikv} is the number of times a_i has value v when $c = k$, and $R^m_{\mathcal{M}_T}$ is the regret:

$$R^m_{\mathcal{M}_T} = \sum_{h_1+\cdots+h_K=N} \sum_{f_{111}+\cdots+f_{11n_1}=h_1} \cdots \sum_{f_{1K1}+\cdots+f_{1Kn_1}=h_K} \cdots \sum_{f_{m11}+\cdots+f_{m1n_m}=h_1} \cdots \sum_{f_{mK1}+\cdots+f_{mKn_m}=h_K} \frac{N!}{h_1!\cdots h_K!}\prod_{k=1}^{K}\left(\frac{h_k}{N}\right)^{h_k} \cdot \prod_{i=1}^{m}\prod_{k=1}^{K}\frac{h_k!}{f_{ik1}!\cdots f_{ikn_i}!}\prod_{v=1}^{n_i}\left(\frac{f_{ikv}}{h_k}\right)^{f_{ikv}}. \quad (8)$$

The trick to make (8) more efficient is to note that we can move all the terms under their respective summation signs, and replace the inner term with the one-dimensional case, which gives

$$R^m_{\mathcal{M}_T} = \sum_{h_1+\cdots+h_K=N} \frac{N!}{h_1!\cdots h_K!}\prod_{k=1}^{K}\left(\frac{h_k}{N}\right)^{h_k} \cdot \prod_{i=1}^{m}\prod_{k=1}^{K} R^1_{h_k,n_i}. \quad (9)$$

This depends only linearly on the number of variables m making it possible to compute (7) for cases with lots of variables provided that the number of value counts are reasonably small. On the other hand, formula (9) is clearly exponential with respect to K. This makes it infeasible for cases like cluster analysis, where typically K can be very big.

It turns out that the recursive formula (5) can also be generalized to the multi-dimensional case. There are, however, cases where even this recursive generalization is too inefficient. One important example is stochastic optimization problems, where typically one must evaluate the cost function thousands or even hundreds of thousands of times. It is clear that for these cases efficient approximations are needed. This will be the subject of the next section.

3 STOCHASTIC COMPLEXITY APPROXIMATIONS

In the previous section we discussed how the NML can be computed efficiently for both one- and multi-dimensional cases. However, we usually had to assume that the variables in our domain do not have too many values. Although the recursive formula (5) is only logarithmic with respect to the number of values, it is still quadratically dependent on the number of data vectors. Therefore, it is necessary to develop approximations to the NML. In this section, we are going to present three such approximations, two of which are well-known (BIC, Rissanen's asymptotic expansion) and a new one based on analytic combinatorics. For each approximation, we instantiate them for both the single multinomial case and the multivariate model class $\mathcal{M}_T$ defined by Equation (6). Furthermore, since we are able to compute the exact NML for these interesting and important cases, it is possible for the first time assess how accurate these approximations really are. This will be the topic of Section 4.

3.1 BAYESIAN INFORMATION CRITERION

The *Bayesian information criterion (BIC)* (Schwarz, 1978; Kass & Raftery, 1994), also known as the Schwarz criterion, is the simplest of the three approximations. For the single multinomial variable case, we get

$$-\log P_{BIC}(\mathbf{x}^N|\mathcal{M}_1) = -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N)) + \frac{K-1}{2}\log(N), \quad (10)$$

where K is the number of values of the multinomial variable. As the name implies, the BIC has a Bayesian interpretation, but it can also be given a formulation in the MDL setting, as showed in (Rissanen, 1989).

In the multi-dimensional case, we easily get

$$-\log P_{BIC}(\mathbf{x}^N|\mathcal{M}_T) = -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N)) + \frac{(K-1)+K\cdot\sum_{i=1}^{m}(n_i-1)}{2}\cdot\log(N). \quad (11)$$

As can be seen, the BIC approximation is very quick to compute and also easy to generalize to more complex model classes. However, it is known that BIC typically favors too simple model classes.

3.2 RISSANEN'S ASYMPTOTIC EXPANSION

As proved in (Rissanen, 1996), for model classes that satisfy certain regularity conditions, an asymptotic expansion can be derived. The most important condition is that the Central Limit Theorem should hold for the maximum likelihood estimators for all the elements in the model class.

The precise regularity conditions can be found in (Rissanen, 1996). The expansion is as follows:

$$-\log P_{RIS}(\mathbf{x}^N|\mathcal{M}) = -\log P(\mathbf{x}^N|\hat{\theta}(\mathbf{x}^N)) + \frac{k}{2}\log\frac{N}{2\,\pi} + \log\int\sqrt{|I(\theta)|}d\theta + o\,(1)\,, \quad (12)$$

where the integral goes over all the possible parameter vectors $\theta \in \mathcal{M}$, and $I(\theta)$ is the (expected) Fisher information matrix. The first term is the familiar negative logarithm of maximum likelihood. The second term measures the complexity that is due to the number of parameters in the model. Finally, the last term measures the complexity that comes from the local geometrical properties of the model space. For a more precise discussion, see (Grünwald, 1998).

Rissanen's asymptotic expansion for a single multinomial variable is discussed in (Rissanen, 1996), and with our notation it is given by

$$-\log P_{RIS}(\mathbf{x}^N|\mathcal{M}_1) = -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N)) + \frac{K-1}{2}\log\left(\frac{N}{2\,\pi}\right) + \log\left(\frac{\pi^{K/2}}{\Gamma\left(\frac{K}{2}\right)}\right) + o\,(1)\,, \quad (13)$$

where $\Gamma(\cdot)$ is the Euler gamma function.

For the multi-dimensional case, we have earlier (Kontkanen et al., 2000) derived the square root of the determinant of the Fisher information for model class $\mathcal{M}_T$:

$$\sqrt{|I(\theta)|} = \prod_{k=1}^{K}\alpha_k^{\frac{1}{2}\left(\sum_{i=1}^{m}(n_i-1)-1\right)}\prod_{i=1}^{m}\prod_{k=1}^{K}\prod_{v=1}^{n_i}\theta_{ikv}^{-\frac{1}{2}}, \quad (14)$$

where $\alpha_k = P(c = k)$ and $\theta_{ikv} = P(a_i = v|c = k)$. To get (12), we need to integrate this expression over the parameters. Fortunately, this is relatively easy since this expression is a product of Dirichlet integrals, yielding

$$\begin{aligned}\int\sqrt{|I(\theta)|}d\theta &= \int\prod_{k=1}^{K}\alpha_k^{\frac{1}{2}\left(\sum_{i=1}^{m}(n_i-1)-1\right)}\cdot\prod_{i=1}^{m}\prod_{k=1}^{K}\prod_{v=1}^{n_i}\theta_{ikv}^{-\frac{1}{2}}d\theta \\ &= \frac{\prod_{k=1}^{K}\Gamma\left(\frac{1}{2}\left(\sum_{i=1}^{m}(n_i-1)+1\right)\right)}{\Gamma\left(\frac{K}{2}\left(\sum_{i=1}^{m}(n_i-1)+1\right)\right)} \cdot\prod_{i=1}^{m}\prod_{k=1}^{K}\frac{\pi^{n_i/2}}{\Gamma\left(\frac{n_i}{2}\right)}, \quad (15)\end{aligned}$$

and after simplifications we get

$$\begin{aligned}-\log P_{RIS}(\mathbf{x}^N|\mathcal{M}_T) &= -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N)) \\ &+ \frac{(K-1)+K\sum_{i=1}^{m}(n_i-1)}{2}\log\left(\frac{N}{2\,\pi}\right) \\ &+ K\cdot\log\Gamma\left(\frac{1}{2}\left(\sum_{i=1}^{m}(n_i-1)+1\right)\right) \\ &- \log\Gamma\left(\frac{K}{2}\left(\sum_{i=1}^{m}(n_i-1)+1\right)\right) \\ &+ K\cdot\sum_{i=1}^{m}\left(\frac{n_i}{2}\log\pi - \log\Gamma\left(\frac{n_i}{2}\right)\right) + o\,(1)\,. \quad (16)\end{aligned}$$

Clearly, Rissanen's asymptotic expansion is efficient to compute, but for more complex model classes than our $\mathcal{M}_T$, the determinant of the Fisher information is no longer a product of Dirichlet integrals, which might cause technical problems.

3.3 SZPANKOWSKI APPROXIMATION

Theorem 8.32 in (Szpankowski, 2001) gives the redundancy rate for memoryless sources. The theorem is based on analytic combinatorics and generating functions, and can be used as a basis for a new NML approximation. Redundancy rate for memoryless sources is actually the regret for a single multinomial variable, and thus we have

$$\begin{aligned}-\log P_{SZP}(\mathbf{x}^N|\mathcal{M}_1) &= -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N)) \\ &+\frac{K-1}{2}\log\left(\frac{N}{2}\right)+\log\left(\frac{\sqrt{\pi}}{\Gamma\left(\frac{K}{2}\right)}\right)+\frac{\sqrt{2}\,\Gamma\left(\frac{K}{2}\right)}{3\sqrt{N}\,\Gamma\left(\frac{K}{2}-1/2\right)} \\ &+\left(\frac{3+K(K-2)(2K+1)}{36}-\frac{\Gamma^2\left(\frac{K}{2}\right)K^2}{9\Gamma^2\left(\frac{K}{2}-1/2\right)}\right)\cdot\frac{1}{N} \\ &+\mathcal{O}\left(\frac{1}{N^{3/2}}\right). \quad (17)\end{aligned}$$

For the multi-dimensional case we can use the factorized form (9) of the exact NML. Let $\hat{R}^1_{K,N}$ denote the regret approximation in (17) with N data vectors and K possible values. Now we can write

$$\begin{aligned}-\log P_{SZP}(\mathbf{x}^N|\mathcal{M}_T) &= -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N)) \\ &+\log\sum_{h_1+\cdots+h_K=N}\left(\frac{N!}{h_1!\cdots h_K!}\prod_{k=1}^{K}\left(\frac{h_k}{N}\right)^{h_k}\right. \\ &\left.\cdot\prod_{i=1}^{m}\prod_{k=1}^{K}\hat{R}^1_{n_i,h_k}\right)+\mathcal{O}\left(\frac{1}{N^{3/2}}\right). \quad (18)\end{aligned}$$

The time complexity of this approximation grows exponentially with K. However, we believe that similar approximation to (17) can be derived for model class $\mathcal{M}_T$ so that this exponentiality could be removed. This is a topic for future work.

4 EMPIRICAL RESULTS

As noted in the previous section, since we are able to compute the exact NML for model classes discussed in this paper, we have a unique opportunity to test how accurate the NML approximations really are. The first thing to notice is that since all three approximations presented contain the maximum likelihood term, we can ignore it in the comparisons and concentrate on the (log-)regret. Notice that since the regret is constant given the model class (i.e., it does not depend on observed data), we avoid the problem of trying to choose representative and unbiased data sets for the experiments.

We conducted two sets of experiments corresponding to the single multinomial case and the multivariate model class $\mathcal{M}_T$. In the following, we will use the following abbreviations for the approximations:

- BIC: Bayesian information criteria presented in Section 3.1.
- RIS: Rissanen's asymptotic expansion presented in Section 3.2.
- SZP: Szpankowski-based approximation presented in Section 3.3.

We start with the one-dimensional case. Figures 1, 2 and 3 show the differences between the three approximations and the exact log-regret as a function of the data size N with a different K, i.e., with a different number of values for the single variable. Cases with $K = 2$, $K = 4$ and $K = 9$ are shown.

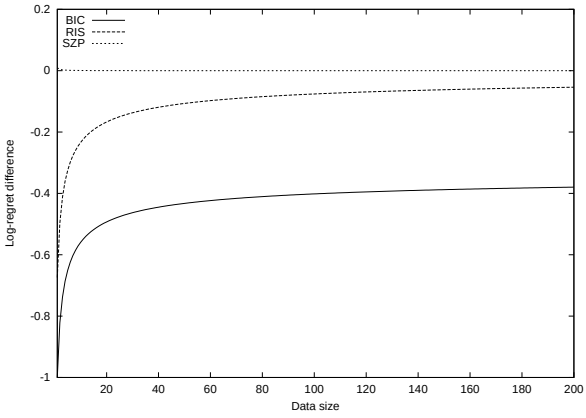

Figure 1: NML approximation results with a single multinomial variable having 2 values.

From these figures we see that the SZP approximation is clearly the best of the three. Furthermore, it is remarkably accurate: just after a few vectors the error is practically zero. The second best approximation is RIS, which takes about 100 data vectors or so to converge to a level near zero.

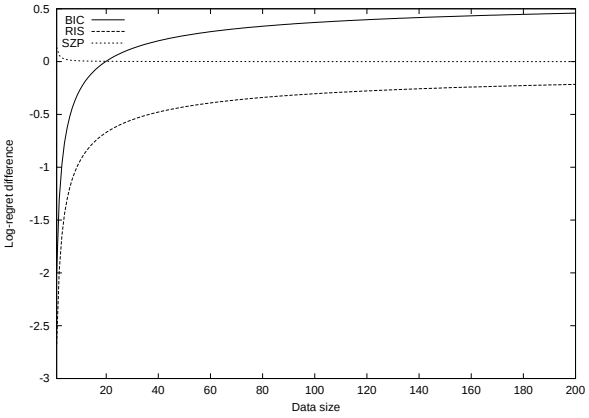

Figure 2: NML approximation results with a single multinomial variable having 4 values.

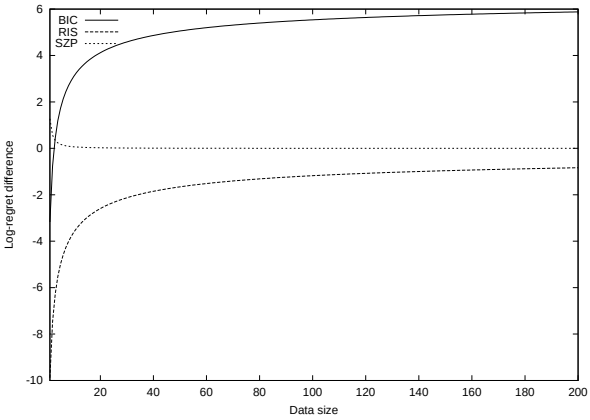

Figure 3: NML approximation results with a single multinomial variable having 9 values.

However, unlike SZP approximation, the convergence of RIS seems to get slower with increasing K. From figures 2 and 3 we see that when the test setting becomes more complex (with $K = 4$ and $K = 9$), BIC starts to overestimate the regret, and thus favors too simple models.

For the multidimensional case we tested with several values for the number of variables m. The results were very similar, so we show here only the case with 30 variables and 2 or 4 values. The special (clustering) variable c was taken to be binary in all tests. The results are shown in Figures 4 and 5.

From the results we can conclude that the SZP approximation is the best and prominently accurate approximation also in the multivariate case. Furthermore, it converged only after few data vectors also in this more complex setting. Rissanen's asymptotic expansion works still reasonably well, but the converge is slower than in the single multinomial case. The BIC approximation overestimates the regret in both cases, and becomes very inaccurate in more complex cases (as can be seen in Figure 5).

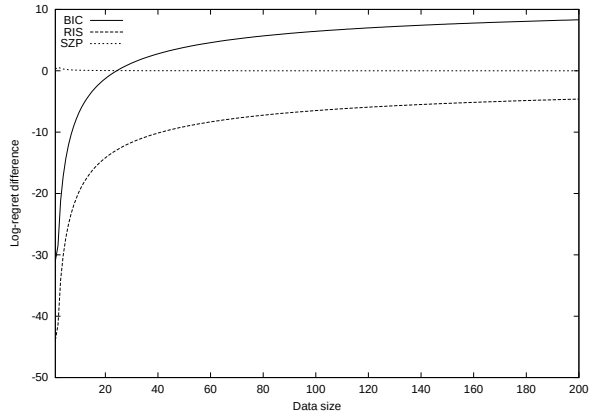

Figure 4: NML approximation results with 30 multinomial variables having 2 values.

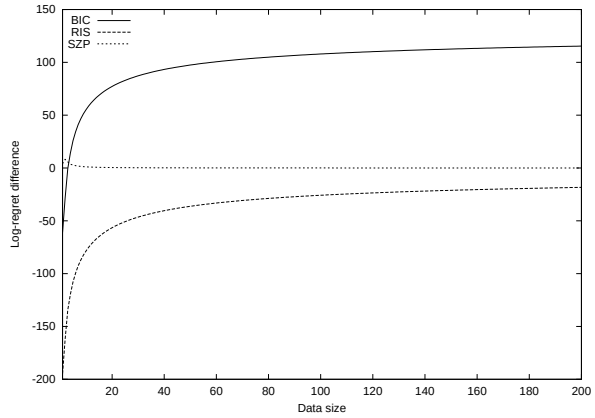

Figure 5: NML approximation results with 30 multinomial variables having 4 values.

5 CONCLUSION AND FUTURE WORK

In this article we have investigated how to compute the stochastic complexity both exactly and approximatively in an attempt to widen the application potential of the MDL principle. We showed that in the case of discrete data the exact form of SC can be computed for several important cases. Particularly interesting was the multi-dimensional model class case, which opens up several application possibilities for the MDL in problems like data clustering.

In addition to exact computation methods, we presented and instantiated three stochastic complexity approximations, and compared their accuracy. The most interesting and important observation was that the new approximation based on analytic combinatorics was significantly better than the older ones. It was also shown to be accurate already with very small sample sizes. Furthermore, the accuracy did not seem to get worse even for the more complex cases. This gives a clear indication that this approximation will also be useful for the cases where exact SC is not efficiently computable.

In the future, on the theoretical side, our goal is to extend the SZP approximation to more complex cases like general graphical models. Secondly, we will research supervised versions of SC, designed for supervised prediction tasks such as classification. On the application side, we have already conducted preliminary tests with MDL clustering by using proprietary real-world industrial data. The preliminary results are very encouraging: according to domain experts we have consulted, the clusterings found with MDL are much better than the ones found with traditional approaches. It is likely that the methods presented here can be used in several other application areas as well with similar success.

Acknowledgments

This research has been supported by the Academy of Finland.

References

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description principle in coding and modeling. *IEEE Transactions on Information Theory*, *44*(6), 2743–2760.

Bernardo, J. (1997). Noninformative priors do not exist. *J. Statist. Planning and Inference*, *65*, 159–189.

Bernardo, J., & Smith, A. (1994). *Bayesian theory.* John Wiley.

Bleistein, N., & Handelsman, R. (1975). *Asymptotic expansions of integrals.* Holt, Rinehart and Winston.

Chickering, D., & Heckerman, D. (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, *29*(2/3), 181–212.

Clarke, B., & Barron, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, *36*(3), 453–471.

Dom, B. (1995). *MDL estimation with small sample sizes including an application to the problem of segmenting binary strings using Bernoulli models* (Tech. Rep. No. RJ 9997 (89085)). IBM Research Division, Almaden Research Center.

Dom, B. (1996). *MDL estimation for small sample sizes and its application to linear regression* (Tech. Rep. No. RJ 10030 (90526)). IBM Research Division, Almaden Research Center.

Grünwald, P. (1998). *The minimum description length principle and reasoning under uncertainty.*

Ph.D. Thesis, CWI, ILLC Dissertation Series 1998-03.

Grünwald, P., Kontkanen, P., Myllymäki, P., Silander, T., & Tirri, H. (1998). Minimum encoding approaches for predictive modeling. In G. Cooper & S. Moral (Eds.), *Proceedings of the 14th international conference on uncertainty in artificial intelligence (UAI'98)* (pp. 183–192). Madison, WI: Morgan Kaufmann Publishers, San Francisco, CA.

Kass, R., & Raftery, A. (1994). *Bayes factors* (Tech. Rep. No. 254). Department of Statistics, University of Washington.

Kontkanen, P., Lahtinen, J., Myllymäki, P., Silander, T., & Tirri, H. (2000). Supervised model-based visualization of high-dimensional data. *Intelligent Data Analysis*, *4*, 213–227.

Kontkanen, P., Myllymäki, P., Silander, T., & Tirri, H. (1998). On Bayesian case matching. In B. Smyth & P. Cunningham (Eds.), *Advances in case-based reasoning, proceedings of the 4th european workshop (EWCBR-98)* (Vol. 1488, pp. 13–24). Springer-Verlag.

Kontkanen, P., Myllymäki, P., Silander, T., & Tirri, H. (1999). On the accuracy of stochastic complexity approximations. In A. Gammerman (Ed.), *Causal models and intelligent data management*. Springer-Verlag.

Kontkanen, P., Myllymäki, P., Silander, T., Tirri, H., & Grünwald, P. (2000). On predictive distributions and Bayesian networks. *Statistics and Computing*, *10*, 39–54.

Kontkanen, P., Myllymäki, P., & Tirri, H. (1996). *Constructing Bayesian finite mixture models by the EM algorithm* (Tech. Rep. No. NC-TR-97-003). ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT).

Merhav, N., & Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory*, *44*(6), 2124–2147.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 445-471.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society*, *49*(3), 223–239 and 252–265.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. New Jersey: World Scientific Publishing Company.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*(1), 40–47.

Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. *Computer Journal*, *42*(4), 260–269.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*(5).

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission*, *23*, 3–17.

Szpankowski, W. (2001). *Average case analysis of algorithms on sequences*. John Wiley & Sons.

Tabus, I., Rissanen, J., & Astola, J. (2002). *Classification and feature gene selection using the normalized maximum likelihood model for discrete regression*. (Unpublished manuscript, Institute of Signal Processing, Tampere University of Technology, Finland)

Xie, Q., & Barron, A. (2000). Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, *46*(2), 431–445.