Pp. 103–117 in *Proceedings of the Causal Models and Statistical Learning Seminar* (London, UK, March 1997).

On the Accuracy of Stochastic Complexity Approximations

Petri Kontkanen, Petri Myllymäki, Tomi Silander, Henry Tirri
Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland
cosco@cs.Helsinki.FI, http://www.cs.Helsinki.FI/research/cosco/

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 determining model complexity, or performing predictive inference. Unfortunately for cases where the data has missing information, computing the stochastic complexity requires marginalizing (integrating) over the missing data, which results even in the discrete data case to computing a sum with an exponential number of terms. Therefore in most cases the stochastic complexity measure has to be approximated. In this paper we will investigate empirically the performance of some of the most common stochastic complexity approximations in an attempt to understand their small sample behavior in the incomplete data framework. In earlier empirical evaluations the problem of not knowing the actual stochastic complexity for incomplete data was circumvented either by using synthetic data, or by comparing the behavior of the stochastic complexity approximation methods to crossvalidated prediction error, approaches which both suffer from validity problems. Our comparison is based on the novel idea of using demonstrably representative small samples from real data sets, and then calculating by "brute force" the exponential sums. This allows for the first time a comparison between the true stochastic complexity and its approximations with real-world data.

1 Introduction

Rissanen [21, 22] has defined the *stochastic complexity* of a data set as the shortest possible code length for the data obtainable by using some fixed set of models. This measure can be used as a tool for solving several difficult problems of great practical importance. For instance, the problem of choosing the proper model complexity (in order to avoid overfitting) can be solved by using the stochastic complexity measure. This type of model selection is common in many machine learning approaches, e.g., in selecting the proper number of hidden units in feed-forward neural networks [11], or in pruning of a decision tree [19].

Stochastic complexity offers also a theoretically solid framework for computing optimal predictive distributions, as will be shown in Section 2. More importantly, for both of these tasks, stochastic complexity can be shown to be an optimal criteria both in information theoretic and Bayesian probability theory frameworks (see the discussion in [2, 22]).

In this paper, we focus on an incomplete data situation, where the sample data contains some missing information. In this case, computing the stochastic complexity requires marginalizing (integrating) over the missing data, which results even in the discrete data case to computing a sum with an exponential number of terms. This is clearly an infeasible task, so in most practical situations stochastic complexity measure has to be approximated.

Although several methods for computing the evidence approximately has been suggested in the literature (see e.g., [1, 3, 4, 14, 22, 25, 27]), the quality of most of these approximations is not well known, except for some asymptotic results. On the other hand, in many real life situations we are typically faced with relatively small data sets. Therefore we will in this paper investigate empirically the performance of some of the most common stochastic complexity approximations in an attempt to understand their small sample behavior in the incomplete data framework. The various stochastic complexity approximation methods used in the experiments are described in more detail in Section 3.

The specific problem domain selected for this empirical study is defined in terms of finite mixture models. This family of models is especially suitable for this purpose, as with finite mixtures we are always faced with missing data created by the basic assumptions defining the model family. Moreover, as demonstrated in Section 4.1, in the finite mixture model family case the stochastic complexity measure can be represented in closed form. However, with the missing data introduced by model assumptions, calculating it in practice requires computing over an exponential sum. It should also be noted that although the finite mixtures are conceptually simple models, our earlier empirical results (see e.g. [28]) show very good performance in predictive inference tasks, when compared to results obtained by more complex model families, such as neural networks or decision trees. In earlier similar studies [20, 24], the model family used has either been too restricted for extending the results to real-world domains, or too general to allow an exact solution to be used for the comparisons.

When trying to evaluate the quality of the stochastic complexity approximations empirically, we encounter the following difficult methodological problem: if calculating the stochastic complexity measure exactly is not feasible for any reasonable sized data set, how do we then evaluate the approximation quality if we do not have any reference measure? In earlier empirical evaluations [5, 17], the problem of not knowing the actual stochastic complexity for incomplete data was circumvented either by using synthetic data, or by comparing the behavior of the stochastic complexity approximation methods to crossvalidated prediction error. However, as pointed out in Section 4.2, using either of these approaches does not necessarily provide correct information about the quality of the approximations.

The key to solving this dilemma lies in an earlier study [15], where we observed that for certain real world data sets we can obtain good predictive models already with very small samples of the full training set. In such cases we do not loose any essential modeling information by replacing the full data set by a small sample. However, for small samples we can actually calculate the exponential sums required for the exact stochastic measure, albeit by using substantial computing power. Therefore we are able to compare the approximations with real data sets to the actual true value of the measure. To our knowledge the comparison presented is first of its nature. The results of the empirical tests performed can be found in Section 4.4.

2 Stochastic complexity and its applications

In the following, let $\mathcal{M}$ denote a set of probability distributions determined by a set of parametric models. In this framework, fixing a specific model, i.e., the parametric form and the specific parameter values, determines a single probability distribution. Consequently, in the following we treat $\mathcal{M}$ as a set of models, instead of as a set of distributions.

Rissanen [21] defined the *stochastic complexity* $SC(\mathcal{D} \mid \mathcal{M})$ of a dataset $\mathcal{D}$ relative to a set of models $\mathcal{M}$ as the shortest code length for $\mathcal{D}$ that can be obtained with the help of models $\mathcal{M}$. In [22], the corresponding code length was defined as

$$SC(\mathcal{D} \mid \mathcal{M}) = -\log P(\mathcal{D} \mid \mathcal{M}) = -\log \int P(\mathcal{D} \mid \Theta, \mathcal{M}) P(\Theta \mid \mathcal{M}) d\Theta, \tag{1}$$

where the integration goes over all the possible models Θ in $\mathcal{M}$. Although Rissanen derived the stochastic complexity measure by using information-theoretic arguments, from (1) we see that the stochastic complexity has a direct link to Bayesian probability theory as the code length is defined with the help of the *marginal likelihood* (or *evidence*) $P(\mathcal{D} \mid \mathcal{M})$. Rissanen has recently [23] introduced an alternative coding scheme for stochastic complexity, which produces for some data sets $\mathcal{D}$ even shorter codes than (1), but in this paper we will focus on the "old" formulation of stochastic complexity.

Stochastic complexity is an interesting measure as it offers solutions to two practically important questions. First of all, in many cases the set $\mathcal{M}$ contains models with a different parametric form. For explorative (data mining) purposes, an important question is which of the model classes (parametric forms) best reflects the probability distribution corresponding to the given sample data $\mathcal{D}$. More precisely, let M_k denote a *model class*, a subset of models each sharing the same parametric form, and let $\mathcal{M}$ be partitioned into K such subsets, $\mathcal{M} = M_1 \cup \ldots \cup M_K$. Now we wish to be able to determine which M_k is best justified by the given data $\mathcal{D}$. In the Bayesian framework, this problem is solved by determining the model class maximizing the posterior probability $P(M_k \mid \mathcal{D})$,

$$P(M_k \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid M_k) P(M_k)}{P(\mathcal{D})}.$$

If all the model classes are assumed to be equally probable a priori, we get

$$P(M_k \mid \mathcal{D}) \propto P(\mathcal{D} \mid M_k) = 2^{-SC(\mathcal{D} \mid M_k)},$$

since $P(\mathcal{D})$ can regarded as a constant. Consequently, the optimal model class can be found by minimizing $SC(\mathcal{D} \mid M_k)$, the stochastic complexity of $\mathcal{D}$ with respect to model class M_k.

The second important application area of stochastic complexity is formed by predictive inference problems, where in the general sense, the task is to compute a predictive distribution for a new data vector $\vec{d}$, given the data $\mathcal{D}$. The set of possible models is here assumed to be restricted to one model class M, which can be determined as described above. A standard approach to this problem is to first determine the model $\hat{\Theta}$ maximizing the posterior probability $P(\Theta \mid \mathcal{D}, M)$ (or the likelihood $P(\mathcal{D} \mid \Theta, M)$), and then to use distribution $P(\vec{d} \mid \hat{\Theta}, M)$ for predictive inference. Nevertheless, from the Bayesian point of view, a more accurate predictive distribution can be obtained by averaging (integrating) over all the possible models in M,

$$\begin{aligned} P(\vec{d} \mid \mathcal{D}, M) &= \int P(\vec{d} \mid \mathcal{D}, \Theta, M) P(\Theta \mid \mathcal{D}, M) d\Theta \\ &= \int \frac{P(\vec{d}, \mathcal{D} \mid \Theta, M) P(\Theta \mid M)}{P(\mathcal{D} \mid M)} d\Theta \\ &\propto \int P(\vec{d}, \mathcal{D} \mid \Theta, M) P(\Theta \mid M) \\ &= P(\vec{d}, \mathcal{D} \mid M) = 2^{-SC(\vec{d}, \mathcal{D} \mid M)}. \end{aligned} \tag{2}$$

Consequently, the Bayes optimal predictive distribution $P(\vec{d} \mid \mathcal{D}, M)$ can be determined if we are able to compute $SC(\vec{d}, \mathcal{D} \mid M)$, the stochastic complexity of the data set $\vec{d} \cup \mathcal{D}$ with respect to model class M.

3 Approximating the stochastic complexity in the incomplete data case

If the model family $\mathcal{M}$ is chosen in such a way that the integral in definition (1) can be computed in feasible time, then the stochastic complexity measure can be used for solving the model class selection and predictive inference problems as described in Section 2. An example of such a simple model family is the Naive Bayes classifier, in which case the model family consists of a single model class (so the model class selection problem disappears), and the predictive distribution (2) can be computed extremely efficiently, as showed in [15]. Nevertheless, in many practical cases the integral is of the form which is not computationally feasible. This situation occurs even with relative simple model families, if the sample data $\mathcal{D}$ is incomplete, i.e., if it contains missing information. In the following we study methods for approximating the stochastic complexity measure in the incomplete data case.

In the sequel, let us use $\mathcal{D}_{\text{obs}}$ for denoting the *observed data* (the sample data given), and $\mathcal{D}_{\text{mis}}$ some missing data which is not given in $\mathcal{D}_{\text{obs}}$. From the Bayesian point of view, the model class selection and predictive inference problems should be solved by using the observed data $\mathcal{D}_{\text{obs}}$ alone by marginalizing out the missing data $\mathcal{D}_{\text{mis}}$. More precisely, the stochastic complexity measure needed for solving these tasks is in this case

$$SC(\mathcal{D}_{\text{obs}} \mid M) = -\log P(\mathcal{D}_{\text{obs}} \mid M) = -\log \int P(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{mis}} \mid M) d\mathcal{D}_{\text{mis}}. \tag{3}$$

However, even if restrict ourselves to the discrete data case, this results to an exponential sum of terms, which makes computing the stochastic complexity (3) an infeasible task. In this paper we compare empirically the four stochastic complexity approximation methods described below.

The *Bayesian information criterion (BIC)* [25, 14], also known as the Schwarz criterion, can be derived by expanding the logarithm of the integrand in (3) around the posterior mode $\hat{\Theta}(\mathcal{D}_{\text{obs}})$,

$$\hat{\Theta}(\mathcal{D}_{\text{obs}}) = \arg\max_{\Theta} P(\Theta \mid \mathcal{D}_{\text{obs}}),$$

which yields

$$SC(\mathcal{D}_{\text{obs}} \mid M) = -\log P(\mathcal{D}_{\text{obs}} \mid M) \approx -\log P(\mathcal{D}_{\text{obs}} \mid \hat{\Theta}(\mathcal{D}_{\text{obs}})) + \frac{1}{2} d \log N, \tag{4}$$

where d is the number of parameters, and N denotes the number of data vectors in $\mathcal{D}_{\text{obs}}$. In the *Akaike information criterion (AIC)* [1], the stochastic complexity is approximated by

$$SC(\mathcal{D}_{\text{obs}} \mid M) = \log P(\mathcal{D}_{\text{obs}}) \approx \log P(\mathcal{D}_{\text{obs}} \mid \hat{\Theta}(\mathcal{D}_{\text{obs}})) - d. \tag{5}$$

The BIC criterion can also been given a formulation in the MDL setting, as showed in [22].

The BIC (and AIC) approximations can also be used as a motivation for another approximation method. For example, from the BIC approximation (4) we know that the probability $P(\mathcal{D}_{\text{obs}} \mid M)$ is approximatively (with increasing N) $P(\mathcal{D}_{\text{obs}} | \hat{\Theta}(\mathcal{D}_{\text{obs}}), M) \cdot C$, where C is a constant depending only on N, and on the dimensionality of $\hat{\Theta}$. Similarly, by using $\tilde{\mathcal{D}}_{\text{mis}}$, an estimate of $\mathcal{D}_{\text{mis}}$, we get

$$P(\mathcal{D}_{\text{obs}}, \tilde{\mathcal{D}}_{\text{mis}} \mid M) \approx P(\mathcal{D}_{\text{obs}}, \tilde{\mathcal{D}}_{\text{mis}} \mid \hat{\Theta}(\mathcal{D}_{\text{obs}}, \tilde{\mathcal{D}}_{\text{mis}}), M) \cdot C.$$

Now by solving C in both cases, and by assuming that $\hat{\Theta}(\mathcal{D}_{\text{obs}}) \approx \hat{\Theta}(\mathcal{D}_{\text{obs}}, \tilde{\mathcal{D}}_{\text{mis}})$, we get

$$\begin{aligned} & SC(\mathcal{D}_{\text{obs}} \mid M) = -\log P(\mathcal{D}_{\text{obs}} \mid M) \\ & \approx -\log \left(P(\mathcal{D}_{\text{obs}} \mid \hat{\Theta}(\mathcal{D}_{\text{obs}}, \tilde{\mathcal{D}}_{\text{mis}}), M) \frac{P(\mathcal{D}_{\text{obs}}, \tilde{\mathcal{D}}_{\text{mis}} \mid M)}{P(\mathcal{D}_{\text{obs}}, \tilde{\mathcal{D}}_{\text{mis}} \mid \hat{\Theta}(\mathcal{D}_{\text{obs}}, \tilde{\mathcal{D}}_{\text{mis}}), M)} \right). \end{aligned} \tag{6}$$

This *Cheeseman-Stutz (CS)* approximation, used in the Autoclass system [4], has in many empirical tests turned out to be quite accurate, yet a computationally efficient approximation of stochastic complexity [5, 17].

In Section 4, the above described stochastic complexity approximation methods (4)–(6) are empirically evaluated by using public domain real world data. The results of these tests inspired us to also experiment with the following simple *local neighborhood (LN)* approximation method, where instead of integrating over all the possible missing data sets, we use only some local neighborhood around $\tilde{\mathcal{D}}_{\text{mis}}$, a locally optimal estimate of the missing data $\mathcal{D}_{\text{mis}}$:

$$\begin{aligned} SC(\mathcal{D}_{\text{obs}} \mid M) &= -\log \int P(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{mis}} \mid M) d\mathcal{D}_{\text{mis}} \\ &\approx -\log \int_{d(\tilde{\mathcal{D}}_{\text{mis}}, \mathcal{D}_{\text{mis}}) < \epsilon} P(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{mis}} \mid M) d\mathcal{D}_{\text{mis}}, \end{aligned} \tag{7}$$

where $d(\tilde{\mathcal{D}}_{\text{mis}}, \mathcal{D}_{\text{mis}})$ is some distance function.

4 Empirical results

4.1 The problem

In the *finite mixture* model family [10, 29], the probability distribution for a data vector $\vec{d}$ is written as a weighted sum of mixture distributions,

$$P(\vec{d}) = \sum_{k=1}^{K} \left(P(Y = y_k) P(\vec{d} | Y = y_k) \right), \tag{8}$$

where Y denotes a latent *clustering random variable*, the values of which are not given in the data $\mathcal{D}_{\mathrm{obs}}$, and K is the number of possible values of Y. In the following, we assume the problem domain to be modeled by m discrete random variables $X_1, \ldots, X_m$. Moreover, the variables $X_1, \ldots, X_m$ are assumed to be independent, given the value of the clustering variable Y, yielding

$$P(\vec{d}) = P(X_1 = x_1, \ldots, X_m = x_m) = \sum_{k=1}^{K} \left(P(Y = y_k) \prod_{i=1}^{m} P(X_i = x_i | Y = y_k) \right).$$

Consequently, the data vectors $\vec{d}, \ldots, \vec{d}_N$ can be thought of being partitioned into K clusters according to the value of the clustering variable Y. The resulting probability distribution can be represented as a simple tree-structured Bayesian network, where the root corresponds to the latent clustering variable Y, and the leaves correspond to the domain variables $X_1, \ldots, X_m$.

It should be noted that by introducing the latent variable Y, with finite mixtures we are always faced with missing data, consisting of values of Y, since by definition, values of a latent variable are never part of the given sample $\mathcal{D}_{\mathrm{obs}}$. Consequently, the finite mixture model family offers a convenient framework for comparing different stochastic complexity approximation methods with missing data. In the sequel, by the *unobserved data* $\mathcal{D}_{\mathrm{mis}}$ we mean a random sample from the distribution of Y, analogous to the *observed data* $\mathcal{D}_{\mathrm{obs}}$, a collection of i.i.d. random samples from the joint distribution of $X_1, \ldots, X_m$.

Both the cluster distribution $P(Y)$ and the intra-class conditional distributions $P_k(X_i)$,

$$P_k(X_i) = P(X_i | Y = y_k),$$

are here assumed to be multinomial. Thus a single finite mixture model can be defined by first fixing K, the model class (the number of the mixing distributions) as described in Section 2, and then by determining the values of the model parameters $\Theta = (\alpha, \Phi), \Theta \in \Omega$, where $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\Phi = (\Phi_{11}, \ldots, \Phi_{1m}, \ldots, \Phi_{K1}, \ldots, \Phi_{Km})$, with the denotations $\alpha_k = P(Y = y_k)$, $\Phi_{ki} = (\phi_{ki1}, \ldots, \phi_{kin_i})$, where $\phi_{kil} = P(X_i = x_{il} | Y = y_k)$.

Since the family of Dirichlet densities is *conjugate* (see e.g. [8]) to the family of multinomials, i.e., the functional form of parameter distribution remains invariant in the prior-to-posterior transformation, we assume that the prior distributions of the parameters are from this family. More precisely, let $(\alpha_1, \ldots, \alpha_K) \sim \mathrm{Di}\,(\mu_1, \ldots, \mu_K)$, and $(\phi_{ki1}, \ldots, \phi_{kin_i}) \sim \mathrm{Di}\,(\sigma_{ki1}, \ldots, \sigma_{kin_i}), k = 1, \ldots, K, i = 1, \ldots, m$, where $\{\mu_k, \sigma_{kil} \mid k = 1, \ldots, K; i = 1, \ldots, m; l = 1, \ldots, n_i\}$ are the *hyperparameters* of the corresponding distributions. Assuming that the parameter vectors α and Φ_{ki} are independent, the joint prior distribution of all the parameters is

$$\mathrm{Di}\,(\mu_1, \ldots, \mu_K) \prod_{k=1}^{K} \prod_{i=1}^{m} \mathrm{Di}\,(\sigma_{ki1}, \ldots, \sigma_{kin_i})$$

As shown in [7, 13], with the above assumptions, the posterior probability of complete data $(\mathcal{D}_{\mathrm{obs}}, \mathcal{D}_{\mathrm{mis}})$ for a K-cluster finite mixture model class M_K can be written as

$$\begin{aligned} P(\mathcal{D}_{\mathrm{obs}}, \mathcal{D}_{\mathrm{mis}} \mid M_K) &= \int P(\mathcal{D}_{\mathrm{obs}}, \mathcal{D}_{\mathrm{mis}} \mid \Theta, M) P(\Theta \mid M_K)\, d\Theta \qquad (9) \\ &= \frac{\Gamma\left(\sum_{k=1}^{K} \mu_k\right)}{\Gamma\left(N + \sum_{k=1}^{K} \mu_k\right)} \prod_{k=1}^{K} \frac{\Gamma(h_k + \mu_k)}{\Gamma(\mu_k)} \prod_{k=1}^{K} \prod_{i=1}^{m} \left(\frac{\Gamma\left(\sum_{l=1}^{n_i} \sigma_{kil}\right)}{\Gamma\left(h_k + \sum_{l=1}^{n_i} \sigma_{kil}\right)} \prod_{l=1}^{n_i} \frac{\Gamma(f_{kil} + \sigma_{kil})}{\Gamma(\sigma_{kil})} \right). \end{aligned}$$

Data set	Size	#Attrs	#Classes
Australian	690	15	2
Glass	214	10	6
Heart disease	270	14	2
Hepatitis	150	20	2
Primary tumor	339	18	21
Breast cancer	286	10	2
Diabetes	768	9	2
Iris	150	5	3
Lymphography	148	19	4

Table 1: The data sets used in our experiments.

As discussed in Section 2, computing the stochastic complexity measure for the incomplete data case requires marginalizing out the missing data $\mathcal{D}_{\text{mis}}$:

$$SC(\mathcal{D}_{\text{obs}} \mid M_K) = -\log P(\mathcal{D}_{\text{obs}} \mid M_K) = -\log \sum_{\mathcal{D}_{\text{mis}}} P(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{mis}} \mid M_K), \qquad (10)$$

where $P(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{mis}} \mid M_K)$ is given by (9), and the sum goes over all the K^N possible clusterings of the data $\mathcal{D}_{\text{obs}} = (\vec{d}_1, \ldots, \vec{d}_N)$ — clearly a computationally infeasible task. In the following section we empirically compare the different stochastic complexity approximation methods described in Section 3 in the task of computing the finite mixture incomplete data stochastic complexity (10).

4.2 The experimental setting

When evaluating the quality of the stochastic complexity approximations we encounter interesting methodological problems. From (10) we saw that due to the exponential summation, calculating the stochastic complexity measure for a finite mixture model class M_K is not feasible for any reasonable sized data set. How do we then evaluate the approximation quality if we do not have any reference measure?

In [5, 17], the problem of not knowing the actual stochastic complexity for incomplete data is circumvented by using synthetic data in a model class selection problem. For synthetic data one assumes that the correct value is "implicitly known", since the number of mixing distributions used for generating data can be controlled. Unfortunately such an empirical study can face serious validity problems. First of all, when data is generated for planned experiments, one does not know whether the results can be generalized to real-world problem domains, or whether they are simply caused by some anomaly in the artificial data generating method. Hence such empirical tests do not necessarily tell us much about the approximation quality for real data sets.

A more severe problem is, however, that when approximative stochastic complexity measures are validated against generated data, one should be extremely careful in providing samples that are representative to the intended mixing distributions. Negative results, i.e., approximations suggesting model classes differing from the "true number" of mixture components M_K can also be caused by the fact that the data in the sample can indeed be described best with a different model class $M_{K'}$, since no finite sample can capture all the information of the generating process. The amounts of data needed to represent the underlying distribution are substantial (thousands of data vectors for parameter spaces of only moderate dimensionality), which defies the whole purpose of finding out the approximation quality for small sample sizes encountered in real life. The results reported in [5, 17] clearly reflect this difficulty.

Since we cannot calculate the stochastic complexity for real data sets, an alternative solution, commonly suggested for model class selection problems, is to compare the stochastic approximation methods against some other, more easily computable measure, such as the leave-one-out crossvalidation measure [26]. In the beginning of this paper we have argued that stochastic complexity provides a principled measure for model selection and prediction tasks. However, we know that the crossvalidation measure is in fact an average value of the last term in the sequential decomposition of the actual stochastic complexity [6], under random re-orderings of the data. Thus it will be very hard to judge the quality of other approximations based on such a coarse measure. The relationship between the stochastic complexity measure (called "scientific criterion") and crossvalidation measure (called "engineering criterion") together with some experimental results in model class selection tasks are discussed in [12].

In an earlier study [15] we have demonstrated that for some commonly used benchmark data sets, on the average very small random samples (less than 10%) are sufficient to construct good predictive models. By good models we mean that they provide prediction performance comparable to the performance of models constructed from the full data set D. This prompted an interesting novel alternative for evaluating the quality of stochastic complexity approximations with real data sets. For such benchmark data sets, without any loss of generality, we can in fact *restrict our comparative study to a small random sample* $D', |D|' << |D|$ *and calculate all the terms in the exponential sum* in (10). Thus by using this "brute force method" we are able to calculate the stochastic complexity exactly, giving us a measure against which we can then compare the approximations for *real data sets*.

For our experiments, we have chosen nine data sets from the collection discussed in [28]. Many of these data sets appeared in the extensive comparative study performed by the StatLog-project [18], and are standard benchmarks in the machine learning community. The main criteria for selecting these particular data sets was the observed learning rate in our earlier experiments, i.e., we preferred data sets for which (on the average) already less than 10% of the full training data was enough to produce good predictive models. Many of the full data sets were less than 270 data vectors, and contained natural data from various problem domains (a short description of the data sets used can be found in Table 1[1]. For our experiments we used random samples D' of size 10 or less, as the need for computing resources increases exponentially with the size of the sample. It should be observed that even for these small samples we used parallel processing power of 30 networked Pentiums running dedicated software on Linux, which allows us use the network as a "supercomputer" for repeated stochastic experiments. Many of the figures presented in Section 4.4 have required more than 30 Pentium CPU days, and are by no means easily obtainable.

[1]The data sets can be obtained from the UCI data repository at URL address "http://www.ics.uci.edu/~mlearn/".

4.3 The algorithms

As noted in Section 4.1, computing the stochastic complexity for a data set $\mathcal{D}$ with N vectors, given a K-cluster finite mixture model class M_K, requires computing over a sum with K^N terms, corresponding to all the possible clusterings of the data. More precisely, let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ denote the possible clusterings (so $n = K^N$). Now we can compute the stochastic complexity of $\mathcal{D}$ by

$$SC(\mathcal{D} \mid M) = -\log \sum_{i=1}^{n} P(\mathbf{Z}_i, \mathcal{D} \mid M), \tag{11}$$

where $P(\mathbf{Z}_i, \mathcal{D} \mid M)$ can be computed by using formula (9).

In the following, we use SC for denoting the (true) stochastic complexity obtained by using formula (11). For approximating this measure, we used in our experiments the following methods:

- BIC: The Bayesian information criterion as defined in (4).
- AIC: The Akaike information criterion as defined in (5).
- C-S: The Cheeseman-Stutz approximation method as defined in (6).
- LN: The local neighborhood approximation method as defined in (7).

In order to be able to use the AIC, BIC,C-S, and LN methods for approximating SC, we need a method for providing estimates of the missing data. In the experiments reported here, we used the *Expectation Maximization (EM)* [9] algorithm for computing $\tilde{\mathcal{D}}_{\text{mis}}$ and $\hat{\Theta}(\mathcal{D}_{\text{obs}}, \tilde{\mathcal{D}}_{\text{mis}} \mid M)$. As EM is a locally optimal search algorithm, the algorithm was repeated in each case 20 times, starting from randomly chosen initial locations. As the final result of the algorithm, we used the model $\hat{\Theta}$ with the highest posterior probability. To see how much the methods depend on the estimates found by the EM algorithm, we also computed their "optimal" performance, i.e., the performance the methods would have obtained if EM would have found the clustering $\hat{\mathbf{Z}}$ maximizing the probability $P(\mathbf{Z}, \mathcal{D} \mid M)$. This type of an optimized version of a method "X" will be denoted by "X*" in the discussion below.

For the LN method, we used in this paper a version where we used formula (11), but instead of summing over all the possible clusterings $\mathbf{Z}$, we used only the minimal local neighborhood of $\hat{\mathbf{Z}}$, i.e., the $N \cdot (K-1)$ clusterings that could be obtained from $\hat{\mathbf{Z}}$ by changing the value of one component only.

4.4 The results

In our first set of experiments, the goal was to study how the exponential sum (11) behaves as a function of the terms included in the sum. For a given data set $\mathcal{D}$, we computed first all the $n = K^N$ possible clusterings by using a network of workstations as described in Section 4.2. These clusterings $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ were then ordered and re-indexed in descending order according to probability $P(\mathbf{Z}, \mathcal{D} \mid M)$. Figure 1 presents the cumulative sum $SC(s)$,

$$SC(s) = -\log \sum_{i=1}^{s} P(\mathbf{Z}_i, \mathcal{D} \mid M), \tag{12}$$

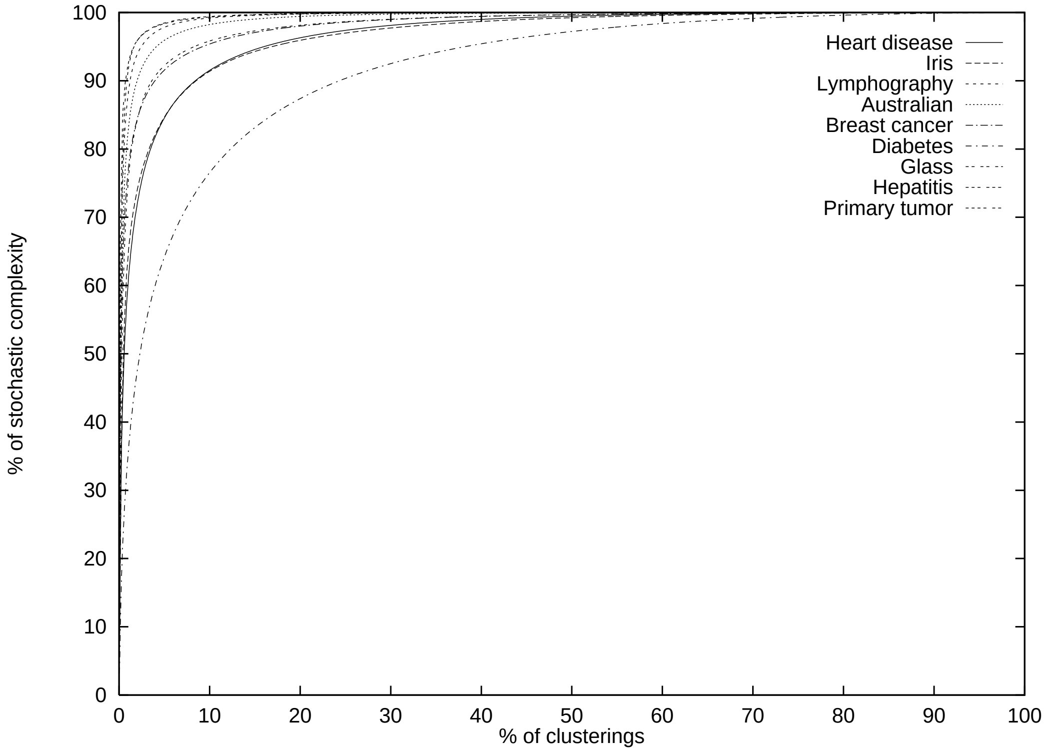

Figure 1: Cumulative distribution of the stochastic complexity for all the data sets, where the terms are arranged in descending order. Here the x-axis is the percentage of the number of terms in the sum (12) and the y-axis gives the percentage of the actual SC measure value obtained by using this number of clusterings.

for the case with $K = 3$ clusters, and for all the data sets in Table 1 with sample sizes of 7 data vectors, and illustrates how the cumulative sum behaves as s approaches n. As the clusterings are arranged in descending order, the "best" clustering can be found on the left. It should be noticed that the behavior of the cumulative sum varies only slightly with the different data sets. For most data sets, less than 10% of the largest terms are needed on the average to reach an error level below 10%. However, one of the data sets (Diabetes) has much less steep curve and needs clearly much more terms (in relative sense) than the others. For this type of data sets we can expect our LN method not to work well.

In Figure 2 we illustrate in the Hepatitis data set case how many terms are needed in the sum (12) for obtaining F% (where F goes from 90 to 99) of the true stochastic complexity as a function of the size of the data set (N). The clusterings are here arranged in descending order as in Figure 1. The results show that although the size of the data space grows exponentially with increasing N, the absolute number of clusterings needed for a good approximation of SC remains surprisingly small. This observation was the initial motivation for our experiments with the local neighborhood approximation approach.

The results for a 10 data vector sample from the Hepatitis data set (with K=3) are presented in Figure 3. Several observations can be made from these results. First, it is clear that the approximations are sensitive to the missing data estimate and behave very poorly with low posterior models. The results also support the observations in [5, 16] where C-S was found to outperform both AIC and BIC. In fact we can also see that

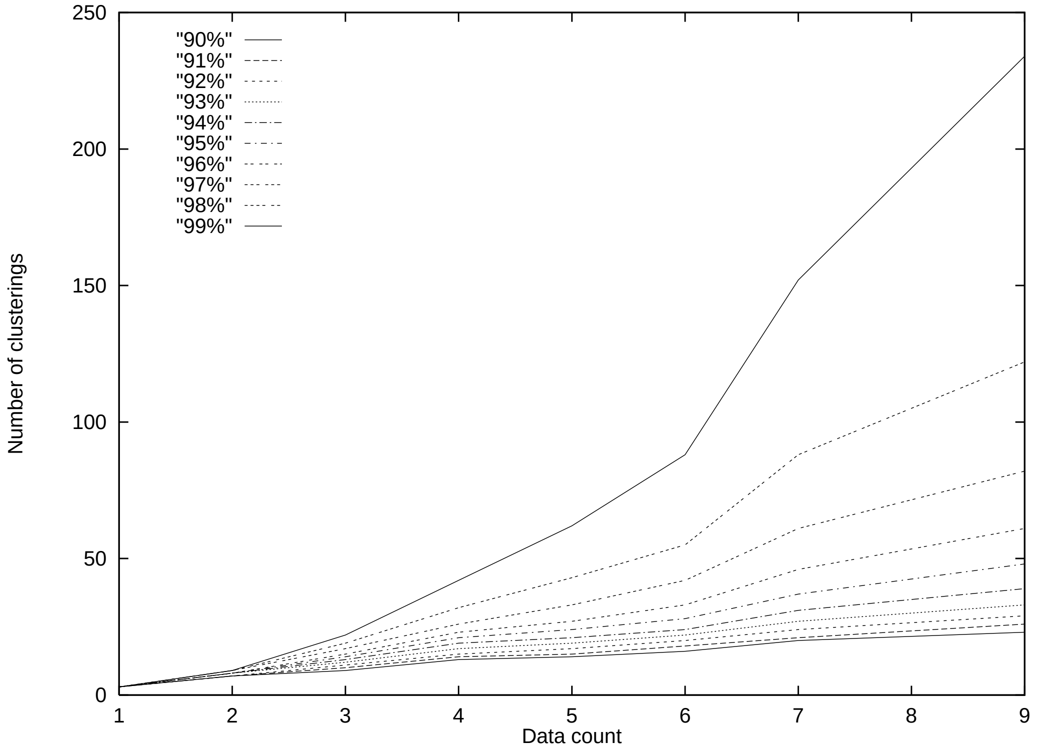

Figure 2: The number of clusterings needed in the sum (12) for obtaining F% (F=90–99) of the true stochastic complexity as a function of the size of the sampled data set.

this does not hold only for the approximations using the estimates found, but also the optimal version C-S* outperforms AIC* and BIC*. However, most interestingly the simple version of the LN approximation outperforms all the other stochastic complexity approximations, both in the "optimal" and "pragmatic" sense.

In the third set of experiments we studied the behavior of the methods in the model class selection setting, which is a typical application of the stochastic complexity measure. Figure 4 illustrates the typical behavior of the methods for a 6 data vector sample from the Hepatitis data set, where the number of mixture components was varied from 1 to 6. Here we again see that the LN method performs best in following the behavior of the correct SC measure, and that the C-S methods outperforms both the BIC and AIC approximations. The behavior of the BIC and AIC methods turned out to be quite similar, which is not surprising considering the minor difference between equations (4) and (5).

5 Conclusion

In this paper we investigated empirically the performance of different stochastic complexity approximation methods in an attempt to understand their small sample behavior for the incomplete data framework. The comparison was based on a novel idea of using small, but demonstrably representative samples from real data sets, which made it possible (although with a considerable computational effort) to compute the stochastic complexity measure exactly by marginalizing out the missing data. This 'brute force" approach allowed us to make a fair comparison between the different approximation methods, since the difference between their results and the correct solution could now

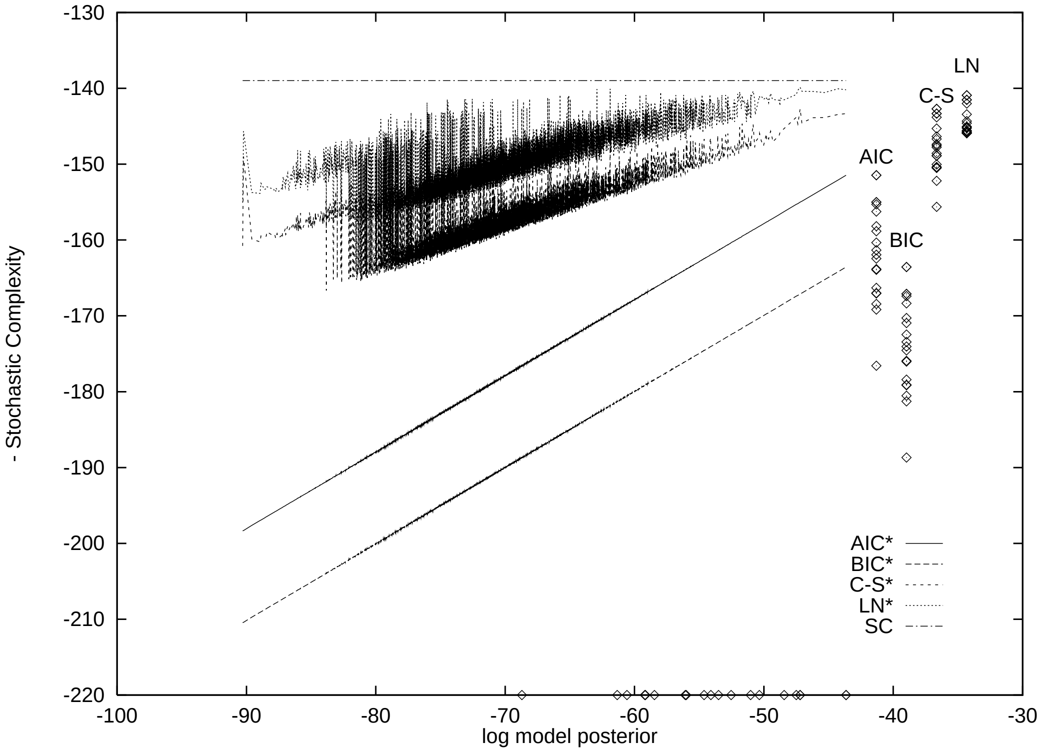

Figure 3: The performance of all the methods for Hepatitis data set (with K=3) with both the optimal estimates and the estimates found by EM. The x-axis is arranged in ascending order of the model posterior. The line in the upper portion of the figure is the actual stochastic complexity. The dots on the x-axis mark the solutions found by the EM algorithm.

be computed. In these experiments, the Cheeseman-Stutz approximation turned out to be superior when compared to the BIC and AIC methods. This supports the results obtained earlier by using alternative approaches for comparing the stochastic complexity approximation methods.

In addition to comparing the performances of the different approximation methods, the experimental setup could be used for exploring the general shape of the stochastic complexity space with incomplete data. The results suggest that the shape of the stochastic complexity space is extremely peaked, so that most of the probability mass is concentrated near few local optima. This observation encouraged us to start experiments with local approximations, where instead of marginalizing out all the possible combinations for the missing data, we concentrated on a small area around a local optimum point. The empirical results show that at least for the small sample cases studied, the local neighborhood approximation performs better than the other approximations that were included in our study. It is also evident that the LN algorithm can be improved substantially by extending the approach to the case where, instead of using only one local neighborhood, we sum over several neighborhoods around local optimum points. We are currently pursuing this line of research further.

Acknowledgements

This research has been supported by the Technology Development Center (TEKES), and by the Academy of Finland. The Primary tumor, the Breast cancer, and the Lympho-

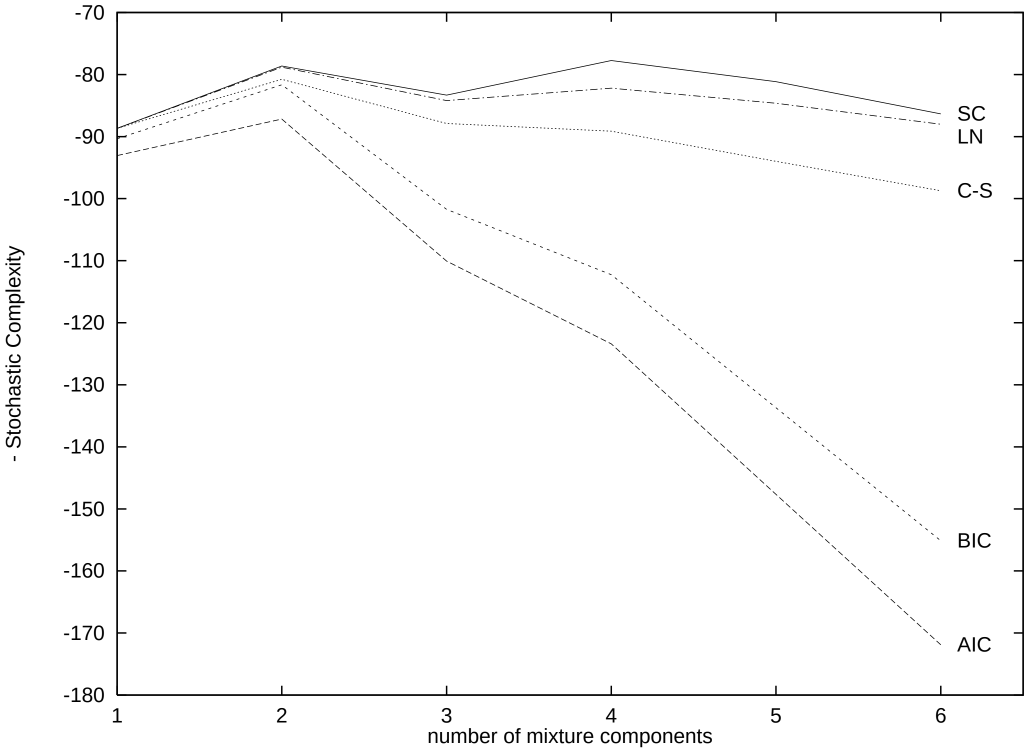

Figure 4: The performance of the approximation criteria for model class selection. Here the x-axis is the number of mixture components, i.e., the model class, and y-axis is the value of the measures (including the stochastic complexity SC).

graphy domains were obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklič for providing the data.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrox and F. Caski, editors, *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kiado.

[2] J.M. Bernardo and A.F.M Smith. *Bayesian theory.* John Wiley, 1994.

[3] H. Bozdogan. On the information-based measure of covariance complexity and its applications to the evaluation of multivariate linear models. *Communications in Statistics - Theory and Methods*, 19(1):221–278, 1990.

[4] P. Cheeseman and J. Stutz. Bayesian classification (AutoClass): Theory and results. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, chapter 6. AAAI Press, Menlo Park, 1996.

[5] D.M. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. In E. Horvitz and F. Jensen, editors, *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, pages 158–168, Portland, Oregon, August 1996. Morgan Kaufmann Publishers.

[6] G. Chow. A comparison of the information and posterior probability criteria for model selection. *Journal of Econometrics*, 16:21–33, 1981.

[7] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[8] M.H. DeGroot. *Optimal statistical decisions.* McGraw-Hill, 1970.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[10] B.S. Everitt and D.J. Hand. *Finite Mixture Distributions.* Chapman and Hall, London, 1981.

[11] S. Haykin. *Neural Networks: A Comprehensive Foundation.* IEEE Press/Macmillan College Publishing Company, New York, 1994.

[12] D. Heckerman and D. Chickering. A comparison of scientific and engineering criteria for bayesian model selection. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, pages 275–281, Ft. Lauderdale, Florida, January 1997.

[13] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, September 1995.

[14] R.E. Kass and A.E. Raftery. Bayes factors. Technical Report 254, Department of Statistics, University of Washington, 1994.

[15] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. Comparing predictive inference methods for discrete domains. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, pages 311–318, Ft. Lauderdale, Florida, January 1997.

[16] P. Kontkanen, P. Myllymäki, and H. Tirri. Comparing Bayesian model class selection criteria by discrete finite mixtures. In D. Dowe, K. Korb, and J. Oliver, editors, *Information, Statistics and Induction in Science*, pages 364–374, Proceedings of the ISIS'96 Conference, Melbourne, Australia, August 1996. World Scientific, Singapore.

[17] P. Kontkanen, P. Myllymäki, and H. Tirri. Experimenting with the Cheeseman-Stutz evidence approximation for predictive modeling and data mining. In *Proceedings of Tenth International FLAIRS Conference (to appear)*, Daytona Beach, Florida, May 1997.

[18] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors. *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, London, 1994.

[19] J.R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.

[20] A. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Technical Report 255, Department of Statistics, University of Washington, 1993.

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

[22] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company, New Jersey, 1989.

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

[24] S. Rosenkranz. *The Bayes factors for model evaluation in hierarchical Poisson model for area counts.* PhD thesis, Department of Biostatistics, University of Washington, 1992.

[25] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[26] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B)*, 36:111–147, 1974.

[27] L. Tierney and J. Kadane. Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Ass.*, 81:82–86, 1986.

[28] H. Tirri, P. Kontkanen, and P. Myllymäki. Probabilistic instance-based learning. In L. Saitta, editor, *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 507–515. Morgan Kaufmann Publishers, 1996.

[29] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York, 1985.