A Fast Normalized Maximum Likelihood Algorithm for Multinomial Data

Petri Kontkanen, Petri Myllymäki
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. In the case of multinomial data, 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. Furthermore, in order to apply NML in practice, one often needs to compute a whole table of these exponential sums. In our previous work, we were able to compute this table by a recursive algorithm. The purpose of this paper is to significantly improve the time complexity of this algorithm. The techniques used here are based on the discrete Fourier transform and the convolution theorem.

1 Introduction

The *Minimum Description Length (MDL)* principle developed by Rissanen [Rissanen, 1978; 1987; 1996] offers a well-founded theoretical formalization of statistical modeling. The main idea of this principle is to represent a set of models (model class) by a single model imitating the behaviour of any model in the class. Such representative models are called *universal*. The universal model itself does not have to belong to the model class as often is the case.

From a computer science viewpoint, the fundamental idea of the MDL principle is *compression of data*. That is, given some sample data, the task is to find a description or *code* of the data such that this description uses less symbols than it takes to describe the data literally. Intuitively speaking, this approach can in principle 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.

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 the MDL is *not* directly related to BIC, but to the formalization described in [Rissanen, 1996]. Unlike Bayesian and many other approaches, the modern MDL principle does not assume that the chosen model class is correct. It even says that there is no such thing as a true model or model class, as acknowledged by many practitioners. The model class is only used as a technical device for constructing an efficient code. For discussions on the theoretical motivations behind the modern definition of the MDL see, e.g., [Rissanen, 1996; Merhav and Feder, 1998; Barron *et al.*, 1998; Grünwald, 1998; Rissanen, 1999; Xie and Barron, 2000; Rissanen, 2001].

The most important notion of the MDL principle is the *Stochastic Complexity (SC)*, which is defined as the shortest description length of a given data relative to a model class $\mathcal{M}$. The modern definition of SC is based on the Normalized Maximum Likelihood (NML) code [Shtarkov, 1987]. Unfortunately, with multinomial data this code involves a sum over all the possible data matrices of certain length. Computing this sum, usually called the *regret*, is obviously exponential. Therefore, practical applications of the NML have been quite rare,

In our previous work [Kontkanen *et al.*, 2003; 2005], we presented a polynomial time (quadratic) method to compute the regret. In this paper we improve our previous results and show how mathematical techniques such as discrete Fourier transform and convolution can be used in regret computation. The idea of applying these techniques for computing a single regret term was first suggested in [Koivisto, 2004], but as discussed in [Kontkanen *et al.*, 2005], in order to apply NML to practical tasks such as clustering, a whole table of regret terms is needed. We will present here an efficient algorithm for this specific task. For a more detailed discussion of this work, see [Kontkanen and Myllymäki, 2005].

2 NML for Multinomial Data

The most important notion of the MDL is the *Stochastic Complexity (SC)*. Intuitively, stochastic complexity is defined as the shortest description length of a given data relative to a model class. To formalize things, let us start with a definition of a model class. Consider a set $\Theta \in \mathbb{R}^d$, where d is a positive integer. A class of parametric distributions indexed by the elements of Θ is called a *model class*. That is, a model

class $\mathcal{M}$ is defined as

$$\mathcal{M} = \{P(\cdot \mid \theta) : \theta \in \Theta\}. \quad (1)$$

Consider now a discrete data set (or matrix) $\mathbf{x}^N = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of N outcomes, where each outcome $\mathbf{x}_j$ is an element of the set $\mathcal{X}$ consisting of all the vectors of the form $(a_1, \ldots, a_m)$, where each variable (or attribute) a_i takes on values $v \in \{1, \ldots, n_i\}$. Given a model class $\mathcal{M}$, the *Normalized Maximum Likelihood (NML)* distribution [Shtarkov, 1987] is defined as

$$P_{NML}(\mathbf{x}^N \mid \mathcal{M}) = \frac{P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N), \mathcal{M})}{\mathcal{R}^N_{\mathcal{M}}}, \quad (2)$$

where $\hat{\theta}(\mathbf{x}^N)$ denotes the *maximum likelihood* estimate of data $\mathbf{x}^N$, and $\mathcal{R}^N_{\mathcal{M}}$ is given by

$$\mathcal{R}^N_{\mathcal{M}} = \sum_{\mathbf{x}^N} P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N), \mathcal{M}), \quad (3)$$

and the sum goes over all the possible data matrices of size N. The term $\mathcal{R}^N_{\mathcal{M}}$ is called the *regret*. The definition (2) is intuitively very appealing: every data matrix is modeled 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.

The stochastic complexity of a data set $\mathbf{x}^N$ with respect to a model class $\mathcal{M}$ can now be defined as the negative logarithm of (2), i.e.,

$$SC(\mathbf{x}^n \mid \mathcal{M}) = -\log \frac{P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N), \mathcal{M})}{\mathcal{R}^N_{\mathcal{M}}} \quad (4)$$

$$= -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N), \mathcal{M}) + \log \mathcal{R}^N_{\mathcal{M}}. \quad (5)$$

As in [Kontkanen *et al.*, 2005], in the sequel we focus on a multi-dimensional model class suitable for cluster analysis. The selected model class has also been successfully applied to mixture modeling [Kontkanen *et al.*, 1996], case-based reasoning [Kontkanen *et al.*, 1998], Naive Bayes classification [Grünwald *et al.*, 1998; Kontkanen *et al.*, 2000b] and data visualization [Kontkanen *et al.*, 2000a].

Let us assume that we have m variables, $(a_1, \ldots, a_m)$, and we also assume the existence of a special variable c (which can be chosen to be one of the variables in our data or it can be latent). Furthermore, given the value of c, the variables $(a_1, \ldots, a_m)$ are assumed to be independent. The resulting model class is denoted by $\mathcal{M}_T$. Suppose the special variable c has K values and each a_i has n_i values. The NML distribution for the model class $\mathcal{M}_T$ is now

$$P_{NML}(\mathbf{x}^N \mid \mathcal{M}_T) = \left[\prod_{k=1}^{K} \left(\frac{h_k}{N}\right)^{h_k} \prod_{i=1}^{m} \prod_{k=1}^{K} \prod_{v=1}^{n_i} \left(\frac{f_{ikv}}{h_k}\right)^{f_{ikv}}\right] \cdot \frac{1}{\mathcal{R}^N_{\mathcal{M}_T,K}}, \quad (6)$$

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 $\mathcal{R}^N_{\mathcal{M}_T,K}$ is the regret term. In [Kontkanen *et al.*, 2005] it was proven that an efficient way to compute the regret term is via the following recursive formula:

$$\mathcal{R}^N_{\mathcal{M}_T,K} = \sum_{r=0}^{N} \frac{N!}{r!(N-r)!} \left(\frac{r}{N}\right)^r \left(\frac{N-r}{N}\right)^{N-r} \cdot \mathcal{R}^r_{\mathcal{M}_T,k_1} \cdot \mathcal{R}^{N-r}_{\mathcal{M}_T,k_2}, \quad (7)$$

where $k_1 + k_2 = K$.

As discussed in [Kontkanen *et al.*, 2005], in order to apply NML to the clustering problem, we need to compute a whole table of regret terms. This table consists of the terms $\mathcal{R}^n_{\mathcal{M}_T,k}$ for $n = 0, \ldots, N$ and $k = 1, \ldots, K$, where K is the maximum number of clusters.

The procedure of computing the regret table starts by filling the first column, i.e., the case $k = 1$, which is trivial (see [Kontkanen *et al.*, 2005]). To compute the column k, for $k = 2, \ldots, K$, the recursive formula (7) can be used by choosing $k_1 = k - 1,\ k_2 = 1$. The time complexity of filling the whole table is $\mathcal{O}\left(K \cdot N^2\right)$. For more details, see [Kontkanen *et al.*, 2005; Kontkanen and Myllymäki, 2005].

In practice, the quadratic dependency on the size of data limits the applicability of NML to small or moderate size data sets. In the next section, we will present a novel, significantly more efficient method for computing the regret table.

3 The Fast NML Algorithm

In this section we will derive a very efficient algorithm for the regret table computation. The new method is based on the Fast Fourier Transform algorithm. As mentioned in the previous section, the calculation of the first column of the regret table is trivial. Therefore, we only need to consider the case of calculating the column k given the first $k-1$ columns. Let us define two sequences $\mathbf{a}$ and $\mathbf{b}$ by

$$a_n = \frac{n^n}{n!}\mathcal{R}^n_{\mathcal{M}_T,k-1}, \quad b_n = \frac{n^n}{n!}\mathcal{R}^n_{\mathcal{M}_T,1}, \quad (8)$$

for $n = 0, \ldots, N$. Evaluating the convolution of $\mathbf{a}$ and $\mathbf{b}$ gives

$$(\mathbf{a} * \mathbf{b})_n = \sum_{h=0}^{n} \frac{h^h}{h!}\mathcal{R}^h_{\mathcal{M}_T,k-1} \frac{(n-h)^{n-h}}{(n-h)!}\mathcal{R}^{n-h}_{\mathcal{M}_T,1} \quad (9)$$

$$= \frac{n^n}{n!} \sum_{h=0}^{n} \frac{n!}{h!(n-h)!} \left(\frac{h}{n}\right)^h \left(\frac{n-h}{n}\right)^{n-h} \cdot \mathcal{R}^h_{\mathcal{M}_T,k-1}\mathcal{R}^{n-h}_{\mathcal{M}_T,1} \quad (10)$$

$$= \frac{n^n}{n!}\mathcal{R}^n_{\mathcal{M}_T,k}, \quad (11)$$

where the last equality follows from the recursion formula (7). This derivation shows that the column k can be computed by first evaluating the convolution (11), and then multiplying each term by $n!/n^n$.

The standard *convolution theorem* states that convolutions can be evaluated via the (discrete) Fourier transform, which in turn can be computed efficiently with the Fast Fourier Transform algorithm (see [Kontkanen and Myllymäki, 2005] for details). It follows that the time complexity of computing the

whole regret table drops to $\mathcal{O}(N \log N \cdot K)$. This is a major improvement over $\mathcal{O}\left(N^2 \cdot K\right)$ obtained by the recursion method of Section 2.

4 Conclusion And Future Work

The main result of this paper was a derivation of a novel algorithm for the regret table computation. The theoretical time complexity of this algorithm allows practical applications of NML in domains with very large datasets. With the earlier quadratic-time algorithms, this was not possible.

In the future, we plan to conduct an extensive set of empirical tests to see how well the theoretical advantage of the new algorithm transfers to practice. On the theoretical side, our goal is to extend the regret table computation to more complex cases like general graphical models. We will also research supervised versions of the stochastic complexity, designed for supervised prediction tasks such as classification.

Acknowledgements

This work was supported in part by the Academy of Finland under the projects Minos and Civi and by the National Technology Agency under the PMMA project. In addition, this work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

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

[Grünwald *et al.*, 1998] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Minimum encoding approaches for predictive modeling. In G. Cooper and S. Moral, editors, *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 183–192, Madison, WI, July 1998. Morgan Kaufmann Publishers, San Francisco, CA.

[Grünwald, 1998] P. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty*. PhD thesis, CWI, ILLC Dissertation Series 1998-03, 1998.

[Koivisto, 2004] M. Koivisto. *Sum-Product Algorithms for the Analysis of Genetic Risks*. PhD thesis, Report A-2004-1, Department of Computer Science, University of Helsinki, 2004.

[Kontkanen and Myllymäki, 2005] P. Kontkanen and P. Myllymäki. Computing the regret table for multinomial data. Technical Report 2005-1, Helsinki Institute for Information Technology (HIIT), 2005.

[Kontkanen *et al.*, 1996] P. Kontkanen, P. Myllymäki, and H. Tirri. Constructing Bayesian finite mixture models by the EM algorithm. Technical Report NC-TR-97-003, ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT), 1996.

[Kontkanen *et al.*, 1998] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On Bayesian case matching. In B. Smyth and P. Cunningham, editors, *Advances in Case-Based Reasoning, Proceedings of the 4th European Workshop (EWCBR-98)*, volume 1488 of *Lecture Notes in Artificial Intelligence*, pages 13–24. Springer-Verlag, 1998.

[Kontkanen *et al.*, 2000a] P. Kontkanen, J. Lahtinen, P. Myllymäki, T. Silander, and H. Tirri. Supervised model-based visualization of high-dimensional data. *Intelligent Data Analysis*, 4:213–227, 2000.

[Kontkanen *et al.*, 2000b] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. On predictive distributions and Bayesian networks. *Statistics and Computing*, 10:39–54, 2000.

[Kontkanen *et al.*, 2003] P. Kontkanen, W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri. Efficient computation of stochastic complexity. In C. Bishop and B. Frey, editors, *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, pages 233–238. Society for Artificial Intelligence and Statistics, 2003.

[Kontkanen *et al.*, 2005] P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri. An MDL framework for data clustering. In P. Grünwald, I.J. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*. The MIT Press, 2005.

[Merhav and Feder, 1998] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, October 1998.

[Rissanen, 1978] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:445–471, 1978.

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

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

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

[Rissanen, 2001] J. Rissanen. Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47(5):1712–1717, July 2001.

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

[Shtarkov, 1987] Yu M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23:3–17, 1987.

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