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)
Finland

http://cosco.hiit.fi/
cosco@hiit.fi

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

CoSCo

Stochastic Complexity

- Central concept in the *Minimum Description Length (MDL)* framework for statistical modeling.
- *Stochastic complexity = the shortest description length of a given data set relative to a model class* $\mathcal{M}$.
- Model class: a set of parametric distributions indexed by elements of $\Theta \in \mathbb{R}^d$:

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

- Old formalizations of SC: BIC, marginal likelihood.
- Modern formalization: **SC = Normalized Maximum Likelihood (NML)**.

Normalized Maximum Likelihood

☞ *Definition*:

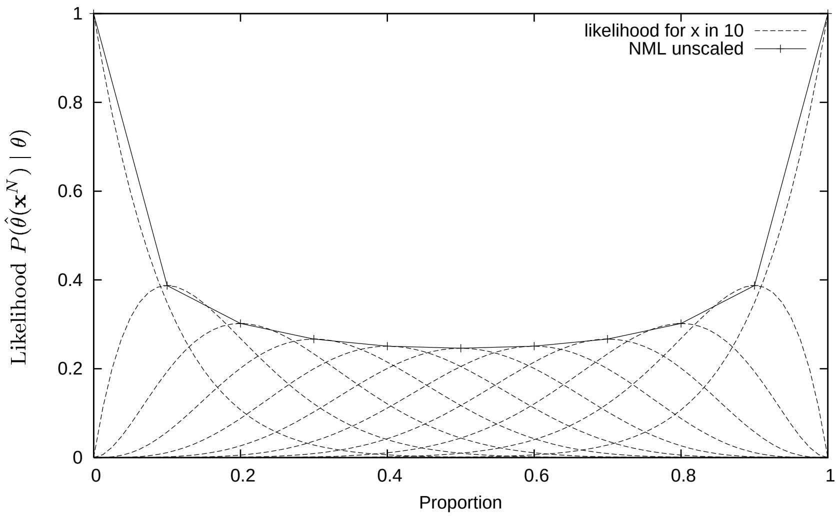
$$P_{NML}(\mathbf{x}^N \mid \mathcal{M}) = \frac{P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N), \mathcal{M})}{\sum_{\mathbf{y}^N} P(\mathbf{y}^N \mid \hat{\theta}(\mathbf{y}^N), \mathcal{M})}, \text{ where} \tag{2}$$

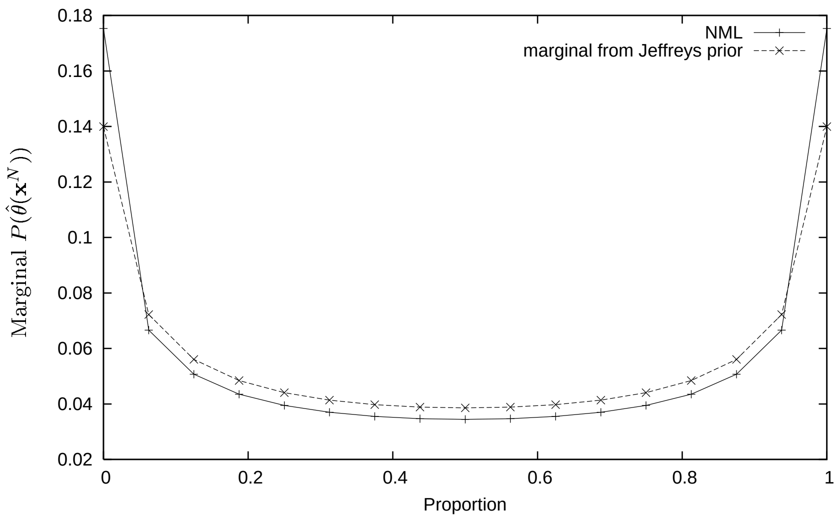
$$\hat{\theta}(\mathbf{x}^n) = \arg\max_{\theta \in \Theta}\{P(\mathbf{x}^n|\theta, \mathcal{M})\}. \tag{3}$$

☞ *Justification*: Distribution (2) is the solution to the following optimization problem:

$$\min_{Q(\mathbf{x}^n)} \max_{\mathbf{x}^n} \log \frac{P(\mathbf{x}^n|\hat{\theta}(\mathbf{x}^n), \mathcal{M})}{Q(\mathbf{x}^n)} \tag{4}$$

NML as a Universal Model





- Gives a *non-informative* distribution over data, <u>not</u> dependent on any explicit prior on $\theta \in \Theta$.
- Asymptotically equivalent to the marginal likelihood with the Jeffreys prior, but different in the finite case.

Exact Computation of NML

- Consider a single multinomial variable $\mathbf{x}$ with K values. The numerator of (2) is easy to compute. The denominator (the *regret*) can be computed by using either *composition summing*:

$$R^1_{K,N} \stackrel{\text{def.}}{=} \sum_{\mathbf{x}^N} P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N), \mathcal{M}_1) = \sum_{h_1+\cdots+h_K=N} \frac{N!}{h_1!\cdots h_K!} \prod_{v=1}^{K} \left(\frac{h_v}{N}\right)^{h_v}. \tag{5}$$

  or *recursion*:

$$R^1_{K,N} = \sum_{h_1+h_2=N} \frac{N!}{h_1!h_2!} \left(\frac{h_1}{N}\right)^{h_1} \left(\frac{h_2}{N}\right)^{h_2} R^1_{k1,h1} R^1_{k2,h2}, \text{where } k_1 + k_2 = K. \tag{6}$$

- Computational complexity is with composition summing $\mathcal{O}(N^{K-1})$ and $\mathcal{O}(N^2 \log K)$ with recursion.

- Extensions to certain multi-dimensional cases possible (See Section 2.4), but computational complexity may become a problem in practice.

Approximations

☞ The Bayesian Information Criterion (BIC):

$$-\log P_{BIC}(\mathbf{x}^N|\mathcal{M}_1) = -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N)) + \frac{K-1}{2}\log(N). \tag{7}$$

☞ The Rissanen asymptotic expansion (RIS):

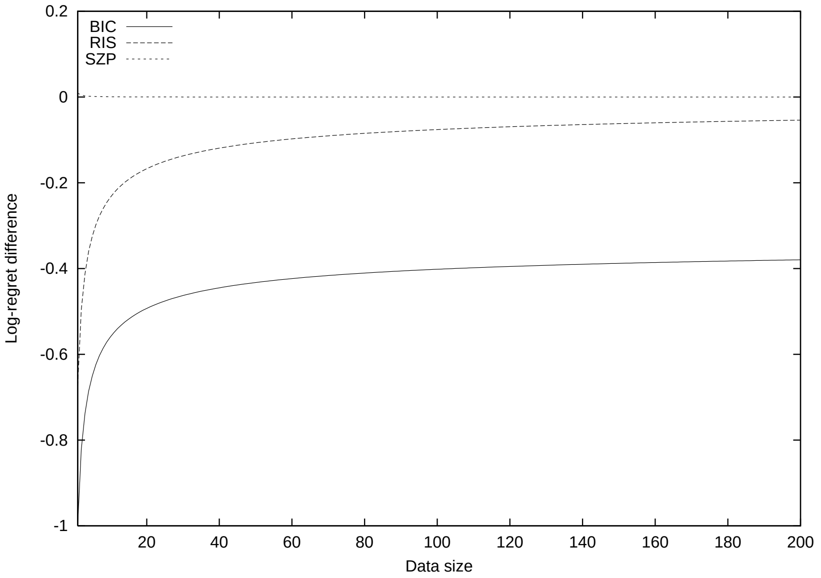
$$-\log P_{RIS}(\mathbf{x}^N|\mathcal{M}_1) = -\log P(\mathbf{x}^N|\hat{\theta}(\mathbf{x}^N)) + \frac{k}{2}\log\frac{N}{2\pi} + \log\int\sqrt{|I(\theta)|}d\theta + o(1) \tag{8}$$

$$= -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N)) + \frac{K-1}{2}\log\left(\frac{N}{2\pi}\right) + \log\left(\frac{\pi^{K/2}}{\Gamma\left(\frac{K}{2}\right)}\right) + o(1)\,. \tag{9}$$

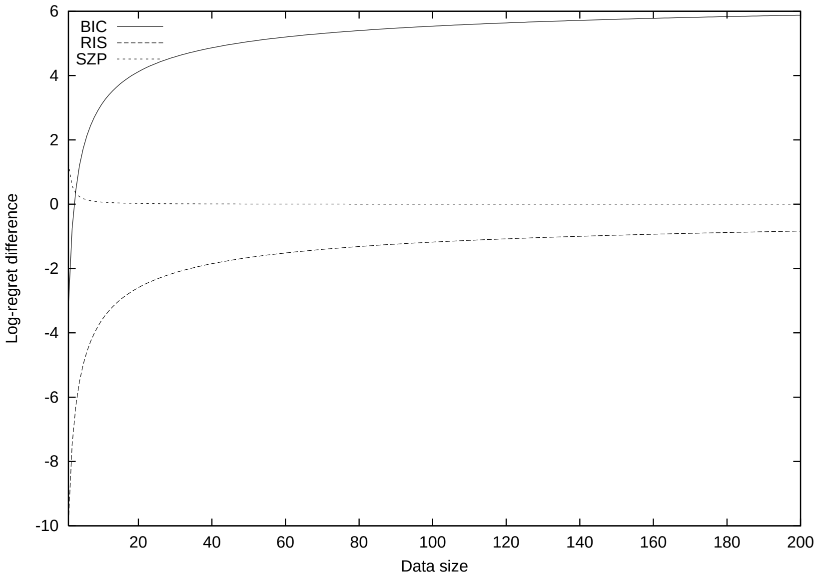
☞ The Szpankowski approximation (SZP):

$$\begin{aligned}-\log P_{SZP}(\mathbf{x}^N|\mathcal{M}_1) &= -\log P(\mathbf{x}^N \mid \hat{\theta}(\mathbf{x}^N)) \\ &+ \frac{K-1}{2}\log\left(\frac{N}{2}\right) + \log\left(\frac{\sqrt{\pi}}{\Gamma\left(\frac{K}{2}\right)}\right) + \frac{\sqrt{2}\,\Gamma\left(\frac{K}{2}\right)}{3\sqrt{N}\,\Gamma\left(\frac{K}{2}-1/2\right)} \\ &+ \left(\frac{3+K(K-2)(2K+1)}{36} - \frac{\Gamma^2\left(\frac{K}{2}\right)K^2}{9\Gamma^2\left(\frac{K}{2}-1/2\right)}\right)\cdot\frac{1}{N} + \mathcal{O}\left(\frac{1}{N^{3/2}}\right).\end{aligned} \tag{10}$$

Empirical Results



☞ Approximation error as a function of sample size with a single multinomial with 2 values.



☞ Approximation error as a function of sample size with a single multinomial with 9 values.

Conclusions

- Exact computation of the NML is feasible for the one-dimensional multinomial case, and for certain multi-dimensional cases.
- For more complex cases, approximations are needed.
- The BIC approximation favors too simple models.
- The Rissanen approximation is reasonable with sufficient amount of data.
- The Szpankowski approximation is remarkably accurate, and for the one-dimensional case it is efficient to compute
- Future work: extending the Szpankowski approximation to multi-dimensional cases in a computationally more efficient manner.