An Empirical Study of Minimum Description Length Model Selection with Infinite Parametric Complexity

Steven de Rooij and Peter Grünwald

February 8, 2006

Abstract

Parametric complexity is a central concept in Minimum Description Length (MDL) model selection. In practice it often turns out to be infinite, even for quite simple models such as the Poisson and Geometric families. In such cases, MDL model selection as based on NML and Bayesian inference based on Jeffreys' prior can not be used. Several ways to resolve this problem have been proposed. We conduct experiments to compare and evaluate their behaviour on small sample sizes.

We find interestingly poor behaviour for the plug-in predictive code; a restricted NML model performs quite well but it is questionable if the results validate its theoretical motivation. A Bayesian marginal distribution with Jeffreys' prior can still be used if one sacrifices the first observation to make a proper posterior; this approach turns out to be most dependable.

1 Introduction

Model selection is the task of choosing one out of a set of alternative hypotheses, or models, for some phenomenon, based on the available data. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two parametric models and D be the observed data. According to the *Minimum Description Length (MDL) Principle* [Rissanen 1989; Barron, Rissanen, and Yu 1998; Grünwald 2005], model selection should proceed by selecting the model that allows for the shortest description of the data. MDL was introduced to a psychological audience in 2000 [Grünwald 2000] and since then, it has been successfully applied in a number of psychological contexts [Myung, Pitt, Zhang, and Balasubramanian 2001; Pitt, Myung, and Zhang 2002; Lee and Navarro 2005; Chater 2005]. A substantial problem when applying MDL in practice is that its optimal implementation based on the *normalised maximum likelihood* (NML) code, does not exist for many (if not most!) standard statistical models. Several remedies have been proposed for this situation. In this paper, we evaluate such alternative procedures empirically for model selection between the simple Poisson and geometric models. We now explain this in more detail.

Modern MDL model selection proceeds by associating so-called *universal codes* (formally defined in Section 2) with each model. We denote the resulting codelength for the data by $L_{\mathcal{M}}(D)$, which can be read as 'the number of bits needed to encode D with the help of model $\mathcal{M}$'. We then pick the model $\mathcal{M}$ that minimises this expression and thus achieves the best compression of the data.

There exist a variety of universal codes, all of which lead to different codelengths and thus to different model selection criteria. While any choice of universal code will lead to an asymptotically 'consistent' model selection method (eventually the right model will be selected), to get good results for small sample sizes it is crucial that we select an efficient code. According to the MDL philosophy, this should be taken to mean a code that compresses the data as

much as possible in a worst-case sense. This is made precise in Section 2. Whenever this code, known as the Shtarkov or NML code, exists, MDL model selection is straightforward to apply and typically leads to good results (see [Grünwald, Myung, and Pitt 2005] and the various experimental papers therein).

However, a worst-case optimal universal code does not exist for many models of practical interest. For example, this holds for models of the form $y = f_\theta(x) + \texttt{noise}$ where $\{f_\theta\}$ is a parameterised set of functions and `noise` is a 0-mean, normally distributed noise term. Such models often appear in a psychological contexts [Myung, Balasubramanian, and Pitt 2000; Navarro 2004]. In such cases it is not clear how MDL model selection should be applied. A variety of remedies to this problem have been proposed. These amount to defining the codelength $L_{\mathcal{M}}(D)$ using a universal code that is not worst-case optimal. In this paper we investigate such remedies empirically for model selection between the Poisson and the geometric model. We chose these two models since they are just about the simplest and easiest-to-analyze models for which the NML code is undefined. We find that some alternative solutions, such as the use of BIC (or, equivalently, in our context, maximum likelihood testing) lead to relatively poor results. Our most surprising finding is the fact that the plug-in code – which has been shown to perform remarkably well in some other contexts [Modha and Masry 1998; Kontkanen, Myllymäki, and Tirri 2001] – shows relatively poor behaviour. We analyze the reasons for this behaviour in Section 6. Since the codes used in these approaches no longer minimise the worst-case regret they are harder to justify theoretically. In fact, as explained in more detail in Section 4.7, the only method that may have an MDL-type justification closely related to that of the original NML code is the Bayesian code with the improper Jeffreys' prior. Perhaps not coincidentally, this also seems the most dependable selection criterion among the ones we tried.

In Section 2 we describe the code that achieves worst-case minimal regret. This code does not exist for the Poisson and geometric distributions. We analyze these models in more detail in Section 3. In Section 4 we describe four different approaches to MDL model selection under such circumstances. We test these criteria by measuring error probability, bias and calibration, as explained in Section 5. The results are evaluated in Section 6. Our conclusions are summarised in Section 7.

2 Universal codes

We define a *binary prefix code* as a one-to-many relation from a set of source symbols (the alphabet) to a set of bit sequences (the code words), such that no code word is a prefix of another. We abbreviate 'binary prefix code' to 'code' from now on. We are only interested in codelengths, not in the code words themselves: there are many codes with the same code word lengths for each source symbol but for our purposes it does not pay to distinguish between them. In what follows we will be concerned with sequences of some number of outcomes: treating the sequence length as a given, we let the source alphabet equal the set of all possible sequences of the given length.

A *universal code* is a code that is universal with respect to a set of codes, in the sense that it codes the data in not many more bits than the best code in the set, *whichever* data sequence is actually observed. We thus define universal codes on an individual sequence basis, as in [Barron et al. 1998], rather than in an expected sense. The difference between the codelength of the universal code and the codelength of the shortest code in the set is called the *regret*, which is a function of a concrete data sequence, unlike "redundancy" which is an expectation value and which we do not use here.

There exists a one-to-one correspondence between codelengths and probability distributions: for any probability distribution, a code can be constructed such that the negative logs of the

probabilities equal the codeword lengths of the outcomes, and vice versa; here we conveniently ignore rounding issues [Grünwald 2005]. Therefore we can phrase our hypothesis testing procedure in terms of *statistical models*, which are sets of probability distributions, rather than sets of codes. In this paper, we define universal codes relative to parametric families of distributions ('models') $\mathcal{M}$, which we think of as sets of distributions or sets of codelength functions, depending on circumstance. Let U be a code with length function L_U. Relative to a given model $\mathcal{M}$ and sample $x^n = x_1, \ldots, x_n$, the *regret* of U is formally defined as

$$L_U(x^n) - [-\ln P(x^n \mid \hat{\theta}(x^n))], \tag{1}$$

where $\hat{\theta}(x^n)$ is the maximum likelihood estimator, indexing the element of $\mathcal{M}$ that assigns maximum probability to the data. It will sometimes be abbreviated to just $\hat{\theta}$. Also note that we compute codelengths in nats rather than bits, this will simplify subsequent equations somewhat.

The correspondence between codes and distributions referred to above amounts to the fact we can transform P_U into a corresponding code, such that the codelengths satisfy, for all x^n in the sample space $\mathcal{X}^n$,

$$L_U(x^n) = -\ln P_U(x^n).$$

Many different constructions of universal codes have been proposed. Some are easy to implement, others have nice theoretical properties. The MDL philosophy [Rissanen 1996; Grünwald 2005] has it that the best universal code minimises the regret (1) in the worst case of all possible data sequences. This "minimax optimal" solution is called the "Normalised Maximum Likelihood" (NML) code, which was first described by Shtarkov, who also observed its minimax optimality properties. The NML-probability of a data sequence for a parametric model is defined as follows:

$$P_{\mathrm{NML}}(x^n) := \frac{P\Big(x^n \mid \hat{\theta}(x^n)\Big)}{\sum_{y^n} P\Big(y^n \mid \hat{\theta}(y^n)\Big)} \tag{2}$$

In this definition the probability of every sequence is proportional to the probability under the ML parameter value; the denominator is just a normalizing factor which is required to get a proper probability distribution. The codelength that corresponds to this probability is called the *stochastic complexity*. Writing $L(x^n) \equiv -\ln P(x^n)$ we get:

$$-\ln P_{\mathrm{NML}}(x^n) = L_{\mathrm{NML}}(x^n) = L\Big(x^n \mid \hat{\theta}\Big) + \ln \sum_{x^n} P\Big(x^n \mid \hat{\theta}(x^n)\Big) \tag{3}$$

The last term in this equation is called the *parametric complexity*. For this particular universal code, the regret (1) is equal to the parametric complexity for all data sequences, and therefore constant across sequences of the same length. It is not hard to deduce that $P_{\mathrm{NML}}(\cdot)$ achieves minimax optimal regret with respect to $\mathcal{M}$. It is usually impossible to compute the parametric complexity analytically, but there exists a good approximation:

$$L_{\mathrm{ANML}}(x^n) := L\Big(x^n \mid \hat{\theta}\Big) + \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int_{\Theta} \sqrt{\det I(\theta)} d\theta \tag{4}$$

Here, n, k and $I(\theta)$ denote the number of outcomes, the number of parameters and the Fisher information matrix respectively. It can be shown that, under regularity conditions on $\mathcal{M}$, we have:

$$\lim_{n \to \infty} \max_{x^n : \hat{\theta}(x^n) \in \Theta} |L_{\mathrm{ANML}}(x^n) - L_{\mathrm{NML}}(x^n)| = 0.$$

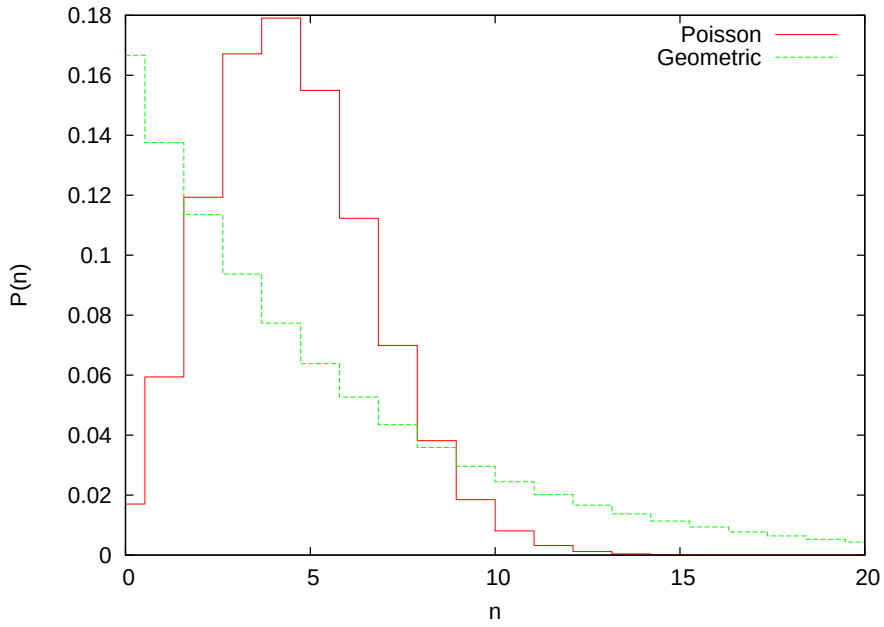

Figure 1: *The Poisson and geometric distributions for* $\mu = 5$.

Since the last term in Equation 4 does not depend on the sample size, it has often been disregarded and many people came to associate MDL only with the first two terms. But the third term can be quite large or even infinite, and it can substantially influence the inference results for small sample sizes.

Interestingly, Equation 4 also describes the asymptotic behaviour of the Bayesian universal code where Jeffreys' prior is used: here MDL and an objective Bayesian method coincide even though their motivation is quite different.

The parametric complexity can be infinite. Many strategies have been proposed to deal with this, but most have a somewhat ad-hoc character. When Rissanen [1996] defines stochastic complexity as (3) he writes that he does so "thereby concluding a decade long search", but as Lanterman [2005] observes, "in the light of these problems we may have to postpone concluding the search just a while longer".

3 The Poisson and geometric models

We investigate MDL model selection between the Poisson and geometric models. Figure 1 may help form an intuition about the probability mass functions of the two distributions. One reason for our choice of models is that they are both single parameter models, so that the dominant $\frac{k}{2} \ln \frac{n}{2\pi}$ term of Equation 4 cancels. This means that at least for large sample sizes, picking the model which best fits the data should always work. We nevertheless observe that for small sample sizes, data which are generated by the geometric distribution are misclassified as Poisson much more frequently than the other way around (see Section 6). So in an informal sense, even though the number of parameters is the same, the Poisson distribution is more prone to 'overfitting'.

To counteract the bias in favor of Poisson that is introduced if we just select the best fitting model, we would like to compute the third term of Equation 4, which now characterises the parametric complexity. But as it turns out, both models have an infinite parametric complexity; the integral in the third term of the approximation also diverges! So in this case it is not

immediately clear how the bias should be removed. This is the second reason why we chose to study the Poisson and geometric models. In Section 4 we describe a number of methods that have been proposed in the literature as ways to deal with infinite parametric complexity; in Section 6 they are evaluated empirically.

Reassuringly, all methods we investigate tend to 'punish' the Poisson model, and thus compensate for this overfitting phenomenon. However, the amount by which the Poisson model is punished depends on the method used, so that different methods give different results.

We parameterise both the Poisson and the Geometric family of distributions by the mean $\mu \in (0, \infty)$, to allow for easy comparison. This is possible because for both models, the empirical mean (average) of the observed data is a sufficient statistic. For Poisson, parameterization by the mean is standard. For geometric, the reparameterisation can be arrived at by noting that in the standard parameterisation, $P(x \mid \theta) = (1-\theta)^x\theta$, the mean is given by $\mu = (1-\theta)/\theta$. As a notational reminder the parameter is called μ henceforth. Conveniently, the ML estimator $\hat{\mu}$ for both distributions is the average of the data (the proof is immediate if we set the derivative to zero).

We will add a subscript P or G to indicate that codelengths are computed with respect to the Poisson model or the geometric model, respectively:

$$L_{\mathrm{P}}(x^n \mid \mu) = -\ln \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{x_i}}{x_i!} = \sum_{i=1}^{n} \ln(x_i!) + n\mu - \ln \mu \sum_{i=1}^{n} x_i \tag{5}$$

$$L_{\mathrm{G}}(x^n \mid \mu) = -\ln \prod_{i=1}^{n} \frac{\mu^{x_i}}{(\mu+1)^{x_i+1}} = n \ln(\mu+1) - \ln\left(\frac{\mu}{\mu+1}\right) \sum_{i=1}^{n} x_i \tag{6}$$

4 Four ways to deal with infinite parametric complexity

In this section we discuss four general ways to deal with the infinite parametric complexity of the Poisson and geometric models when the goal is to do model selection. Each of these four leads to one or sometimes more concrete selection criteria which we put into practice and evaluate in Section 6.

4.1 BIC/ML

One way to deal with the diverging term of the approximation is to just ignore it. The model selection criterion that results corresponds to only a very rough approximation of any real universal code, but it has been used and studied extensively. It was first derived by Jeffreys [1961] as an approximation to the Bayesian marginal likelihood, but it became well-known only when it was proposed by Rissanen [1978] and Schwarz [1978]. While Schwarz gave the same type of derivation as Jeffreys, Rissanen arrived at it in a quite different manner, as an approximation to a two-part codelength. We note that Rissanen already abandoned the idea in the mid 1980's in favour of more sophisticated codelength approximations. Because of its connection to the Bayesian marginal likelihood, it is best known as the BIC (Bayesian Information Criterion).

$$L_{\mathrm{BIC}}(x^n) = L(x^n \mid \hat{\mu}) + \frac{k}{2} \ln n$$

Comparing the BIC to the approximated NML codelength we find that in addition to the diverging term, a $\frac{k}{2} \ln \frac{1}{2\pi}$ term has also been dropped. This curious difference can be safely ignored in our setup, where k is equal to one for both models so the whole term cancels anyway. According to BIC, we must select the geometric model if:

$$0 < L_{\mathrm{P,BIC}}(x^n) - L_{\mathrm{G,BIC}}(x^n) = L_{\mathrm{P}}(x^n \mid \hat{\mu}) - L_{\mathrm{G}}(x^n \mid \hat{\mu})$$

We are left with a generalised likelihood ratio test (GLRT). Such a test the ratio of the probabilities under the two models, $P_{\mathrm{P}}(x^n \mid \hat{\mu}) / P_{\mathrm{G}}(x^n \mid \hat{\mu})$ is compared against a fixed constant η; the BIC criterion thus reduces to a GLRT with $\eta = 0$, which is also known as maximum likelihood (ML) testing. It should be expected that this leads to overfitting and therefore to a bias in favour of the 'more complex' Poisson model. (On a side note, this equivalence of BIC and ML occurs when all models under consideration have the same numbers of parameters; if this is not the case, then BIC may or may not overfit and it usually gives better results than ML.)

4.2 Restricted ANML

One often used method of rescuing the NML approach to MDL model selection is to restrict the range of values that the parameters can take to ensure that the third term of Equation 4 stays finite.

To compute the approximated parametric complexity of the restricted models we need to establish the Fisher information first. Using the formula $I(\theta) = -E_\theta[\frac{d^2}{d\theta^2} L(x \mid \theta)]$, we get:

$$I_{\mathrm{P}}(\mu) = -E_\mu\left[-\frac{x}{\mu^2}\right] = \frac{1}{\mu} \tag{7}$$

$$I_{\mathrm{G}}(\mu) = -E_\mu\left[-\frac{x}{\mu^2} + \frac{x+1}{(\mu+1)^2}\right] = \frac{1}{\mu(\mu+1)} \tag{8}$$

Now we can compute the last term in the parametric complexity approximation (4):

$$\ln \int_0^{\mu^*} \sqrt{I_{\mathrm{P}}(\mu)}\, d\mu = \ln \int_0^{\mu^*} \mu^{-\frac{1}{2}} d\mu = \ln\left(2\sqrt{\mu^*}\right) \tag{9}$$

$$\begin{aligned} \ln \int_0^{\mu^*} \sqrt{I_{\mathrm{G}}(\mu)}\, d\mu &= \ln \int_0^{\mu^*} \frac{1}{\sqrt{\mu(\mu+1)}} d\mu \\ &= \ln\left\{2 \ln\left(\sqrt{\mu^*} + \sqrt{\mu^*+1}\right)\right\}. \end{aligned} \tag{10}$$

The parametric complexities of the restricted models with parameter ranges $(0, \mu^*)$ are both monotonically increasing functions of μ^*. However, the parametric complexity of the restricted Poisson model grows *faster* with μ^* than the parametric complexity of the geometric model, indicating that the Poisson model has more descriptive power, even though the models have the same number of parameters. Let the function $\delta(\mu^*)$ measure the difference between the parametric complexities. Interestingly, it is not hard to prove that this function is still monotonically increasing in μ^* and that grows unboundedly in μ^*.

4.3 Basic restricted ANML

We have experimented with restricted models where the parameter range was restricted to $(0, \mu^*)$ for $\mu^* \in \{10, 100, 1000\}$.

This means that we obtain a model selection criterion that selects the geometric model iff $0 < L_{\mathrm{P,ANML}(\mu^*)}(x^n) - L_{\mathrm{G,ANML}(\mu^*)}(x^n) = L_{\mathrm{P}}(x^n \mid \hat{\mu}) - L_{\mathrm{G}}(x^n \mid \hat{\mu}) + \delta(\mu^*)$. This is equivalent to a GLRT with threshold $\delta(\mu^*)$. For $\mu^* \downarrow 0$ we obtain the BIC/ML selection criterion; higher values of μ^* translate to a selection threshold more in favour of the geometric model.

An obvious conceptual problem with the resulting code is that the imposed restriction is quite arbitrary and requires a priori knowledge about the generating process. But the parameter

range can be interpreted as a hyper-parameter, which can be incorporated into the code using several techniques, some of which we will discuss.

4.4 Two-part ANML

The most obvious way to generalise the restricted ANML codelength is to first encode a suitable parameter range, and then encode the rest of the data using restricted ANML on that range. To do this we need to choose some discretisation, such that whatever $\hat{\mu}$ is, it does not cost many bits to specify an interval that contains it. For a sequence with ML parameter $\hat{\mu}$, we choose to encode the integer $b = \lceil \log_2 \hat{\mu} \rceil$. A decoder, upon reception of such a number b, now knows that the ML parameter value must lie in the range $(2^{b-1}, 2^b]$ (for otherwise another value of b would have been transmitted). So now the data can be transmitted by first sending b and then sending the data using restricted ANML on that range. By taking the logarithm we ensure that the number of bits used in coding the parameter range grows at a negligible rate compared to the codelength of the data itself, but we admit that the code for the parameter range admits of much greater sophistication. We do not really have reason to assume that the best discretisation should be the same for the Poisson and geometric models for example.

The two-part code is slightly redundant, since code words are assigned to data sequences of which the ML estimator lies outside the range that was encoded in the first part – such data sequences cannot occur, since for such a sequence we would have encoded a different range. Furthermore, the two-part code is no longer minimax optimal, so it is no longer clear why it should be better than other universal codes which are not minimax optimal. However, as argued in [Grünwald 2005], whenever the minimax optimal code is not defined, we should aim for a code L which is 'close' to minimax optimal in the sense that, for any compact subset $\mathcal{M}'$ of the parameter space, the additional regret of L on top of the NML code for $\mathcal{M}'$ should be small, e.g. $O(\log \log n)$. The two-part ANML code is one of many universal codes satisfying this 'almost minimax optimality'. While it may not be better than another almost minimax optimal universal code, it certainly is better than universal codes which do not have the almost minimax optimality property.

4.5 Renormalised Maximum Likelihood

Related to the two-part restricted ANML, but more elegant, is Rissanen's *renormalised maximum likelihood* (RNML) code, [Rissanen 2000; Grünwald 2005]. This is perhaps the most widely known approach to deal with infinite parametric complexity. The idea here is that the NML distribution is well-defined *if* the parameter range is restricted to, say, the range $(0, \mu^*)$. Letting P_{NML,μ^*} be the NML distribution relative to this restricted model, we can now define a new parametric model, with μ^* as the parameter and the corresponding restricted NML distributions P_{NML,μ^*} as elements. For this new model we can again compute the NML distribution! To do this, we need to compute the ML value for μ^*, which in this case can be seen to be as small as possible such that $\hat{\mu}$ still falls within the range, in other words, $\mu^* = \hat{\mu}$.

If this still leads to infinite parametric complexity, we define a hyper-hyper-parameter. We repeat the procedure until the resulting complexity is finite. Unfortunately, in our case, after the first renormalisation, both parametric complexities are still infinite; we did not manage to perform a second renormalisation. Therefore, we have not experimented with the RNML code.

4.6 Plug-in predictive code

The *plug-in predictive code*, or prequential ML code, is an attractive universal code because it is usually a lot easier to implement than either NML or a Bayesian code. Moreover, its implemen-

tation hardly requires any arbitrary decisions. Here the outcomes are coded sequentially using the probability distribution indexed by the ML estimator for the previous outcomes [Dawid 1984; Rissanen 1984]; for a general introduction see [Wagenmakers, Grünwald, and Steyvers 2006] or [Grünwald 2005].

$$L_{\text{PIPC}}(x^n) = \sum_{i=1}^{n} L\big(x_i \mid \hat{\mu}(x^{i-1})\big)\,,$$

where $L\big(x_i \mid \hat{\mu}(x^{i-1})\big) = -\ln P\big(x_i \mid \hat{\mu}(x^{i-1})\big)$ is the number of nats needed to encode outcome x_i using the code based on the ML estimator on x^{i-1}. We further discuss the motivation for this code in Section 6.1.

For both the Poisson model and the geometric model, the maximum likelihood estimator is not well-defined until after a nonzero outcome has been observed (since 0 is not inside the allowed parameter range). This means that we need to use another code for the first few outcomes. It is not really clear how we can use the model assumption (Poisson or geometric) here, so we pick a simple code for the nonnegative integers that does not depend on the model. This will result in the same codelength for both models; therefore it does not influence which model is selected. Since there are usually only very few initial zero outcomes, we may reasonably hope that the results are not too distorted by our way of handling this startup problem. We note that this startup problem is an inherent feature of the predictive plug-in approach [Dawid 1984; Rissanen 1989], and our way of handling it is in line with the suggestions in [Dawid 1984].

4.7 Objective Bayesian approach

In the Bayesian framework we select a prior $w(\theta)$ on the unknown parameter and compute the marginal likelihood

$$P_{\text{BAYES}}(x^n) = \int_{\Theta} P(x^n \mid \theta)\, w(\theta) d\theta, \tag{11}$$

which corresponds to a universal code $L_{\text{BAYES}}(x^n) = -\ln P_{\text{BAYES}}(x^n)$. Like the NML, this can be approximated with an asymptotic formula. For exponential families such as the models under consideration, we have [Balasubramanian 1997]:

$$L_{\text{ABAYES}}(x^n) := L\Big(x^n \mid \hat{\theta}\Big) + \frac{k}{2} \ln \frac{n}{2\pi} + \ln \frac{\sqrt{\det I(\theta)}}{w(\theta)}, \tag{12}$$

where the asymptotic behaviour is the same as for the approximation of the NML codelength, roughly $L_{\text{ABAYES}}(x^n) - L_{\text{BAYES}}(x^n) \to 0$ as $n \to \infty$ (see below Eq. (4) for details). Objective Bayesian reasoning suggests we use Jeffreys' prior for several reasons; one reason is that it is uniform over all 'distinguishable' elements of the model [Balasubramanian 1997], which implies that the obtained results are independent of the parametrisation of the model [Jeffreys 1961]. It is defined as follows:

$$w(\theta) = \frac{\sqrt{\det I(\theta)}}{\int_{\Theta} \sqrt{\det I(\theta)} d\theta}. \tag{13}$$

Unfortunately, the normalisation factor in Jeffreys' prior diverges for both the Poisson model and the geometric model. But if one is willing to accept a so-called *improper* prior, which is not normalised, then it is possible to compute a perfectly proper Bayesian posterior, after observing the first outcome, and use that as a prior to compute the marginal likelihood of the rest of the data. Refer to [Bernardo and Smith 1994] for more information on objective Bayesian theory. The resulting universal codes with lengths $L_{\text{BAYES}}(x_2, \ldots, x_n \mid x_1)$ are, in fact, *conditional* on the first outcome. Recent work by [Liang and Barron 2005] suggests that, at least asymptotically and for one-parameter models, the universal code achieving the minimal *expected redundancy*

conditioned on the first outcome is given by the Bayesian universal code with the improper Jeffreys' prior. Li and Barron only prove this for scale and location models, but their result at least suggests that the same would still hold for general exponential families such as Poisson and geometric. It is possible to define MDL inference in terms of either the expected redundancy or of the worst-case regret. In fact, the resulting procedures are very similar, see [Barron et al. 1998]. Thus, we have a tentative justification for using Jeffreys' prior also from an MDL point of view, on top of its justification in terms of objective Bayes.

It can be argued that using the first outcome for conditioning rather than some other outcome is arbitrary while it does influence the results. On the other hand, the posterior after having observed all data will be the same whatever outcome is elected to be the special one that we refrain from encoding. It also seems preferable to let results depend on arbitrary properties of the data than to let it depend on arbitrary decisions of the scientist, such as the choice for a maximum value for μ^* in the case of the restricted ANML criterion.

As advocated for instance in [Berger and Pericchi 1997], arbitrariness can be reduced by conditioning on every outcome in turn and then using the mean or median codelength one so obtains. We have not gone to such lengths in this study.

We compute Jeffreys' posterior after observing one outcome, and use it to find the Bayesian marginal likelihoods. We write x_i^j to denote $x_i, \ldots, x_j$ and $\hat{\mu}(x_i^j)$ to indicate which outcomes determine the ML estimator, finally we abbreviate $s_n = x_1 + \ldots + x_n$. The goal is to compute $P_{\text{BAYES}}(x_2^n \mid x_1)$ for the Poisson and geometric models. As before, the difference between the corresponding codelengths defines a model selection criterion. We also compute $P_{\text{ABAYES}}(x_2^n \mid x_1)$ for both models, the approximated version of the same quantity, based on approximation formula (12). Equations for the Poisson and geometric models are presented below.

Bayesian code for the Poisson model We compute Jeffreys' improper prior and the posterior after observing one outcome:

$$w_{\text{P}}(\mu) \propto \sqrt{I_{\text{P}}(\mu)} = \mu^{-\frac{1}{2}}; \tag{14}$$

$$w_{\text{P}}(\mu \mid x_1) = \frac{P_{\text{P}}(x_1 \mid \mu)\, w_{\text{P}}(\mu)}{\int_0^\infty P_{\text{P}}(x_1 \mid \theta)\, w_{\text{P}}(\theta)\, d\theta} = \frac{e^{-\mu}\mu^{x_1-\frac{1}{2}}}{\Gamma(x_1+\frac{1}{2})}. \tag{15}$$

From this we can derive the marginal likelihood of the rest of the data. The details of the computation are omitted for brevity.

$$\begin{aligned} P_{\text{P,BAYES}}(x_2^n \mid x_1) &= \int_0^\infty P_{\text{P}}(x_2^n \mid \mu)\, w_{\text{P}}(\mu \mid x_1)\, d\mu \\ &= \frac{\Gamma(s_n+\frac{1}{2})}{\Gamma(x_1+\frac{1}{2})} \Big/ \left(n^{s+\frac{1}{2}} \prod_{i=2}^n x_i! \right). \end{aligned} \tag{16}$$

We also compute the approximation for the Poisson model using (12):

$$\begin{aligned} L_{\text{P,ABAYES}}(x_2^n \mid x_1) &= L_{\text{P}}(x_2^n \mid \hat{\mu}(x_2^n)) + \tfrac{1}{2}\ln\frac{n}{2\pi} + \ln\frac{\sqrt{I_{\text{P}}(\hat{\mu}(x_2^n))}}{w_{\text{P}}(\hat{\mu}(x_2^n) \mid x_1)} \\ &= L_{\text{P}}(x_2^n \mid \hat{\mu}(x_2^n)) + \tfrac{1}{2}\ln\frac{n}{2\pi} + \hat{\mu}(x_2^n) - x_1 \ln \hat{\mu}(x_2^n) + \ln\Gamma(x_1+\tfrac{1}{2}). \end{aligned} \tag{17}$$

Bayesian code for the geometric model We perform the same computations for the geometric model. Here we get:

$$w_{\mathrm{G}}(\mu) \propto \mu^{-\frac{1}{2}}(\mu+1)^{-\frac{1}{2}}; \tag{18}$$

$$w_{\mathrm{G}}(\mu \mid x_1) = (x_1+\tfrac{1}{2})\mu^{x_1-\frac{1}{2}}(\mu+1)^{-x_1-\frac{3}{2}}; \tag{19}$$

$$P_{\mathrm{G,BAYES}}(x^n) = (x_1+\tfrac{1}{2})\frac{\Gamma(s+\frac{1}{2})\Gamma(n)}{\Gamma(n+s+\frac{1}{2})}. \tag{20}$$

For the approximation we obtain:

$$L_{\mathrm{G,ABAYES}}(x_2^n \mid x_1) = L_{\mathrm{G}}(x_2^n \mid \hat{\mu}(x_2^n)) + \frac{1}{2}\ln\frac{n}{2\pi} + x_1 \ln\left(1+\frac{1}{\hat{\mu}(x_2^n)}\right) + \tfrac{1}{2}\ln(\hat{\mu}(x_2^n)) - \ln(x_1+\tfrac{1}{2}). \tag{21}$$

5 Experiments

We have now described how to compute or approximate the length of a number of different universal codes, which can be used in an MDL model selection framework. The MDL principle tells us to select the model using which we can achieve the shortest codelength for the data. This coincides with the Bayesian maximum a-posteriori (MAP) model with a uniform prior on the models. In this way each method for computing or approximating universal codelengths defines a model selection criterion, which we want to compare empirically.

Known μ criterion In addition to the criteria that are based on universal codes, as developed in Section 4, we define one additional, 'ideal' criterion to serve as a reference by which the others can be evaluated. The *known* μ criterion cheats a little bit: it computes the codelength for the data with knowledge of the mean of the generating distribution. If the mean is μ, then the known μ criterion selects the Poisson model if $L_{\mathrm{P}}(x^n \mid \mu) < L_{\mathrm{G}}(x^n \mid \mu)$. Since this criterion uses extra knowledge about the data, it should be expected to perform better than the other criteria. The theoretical analysis of the known μ criterion is helped by the circumstance that (1) one of the two hypotheses equals the generating distribution and (2) the sample consists of outcomes which are i.i.d. according to this distribution. In [Cover and Thomas 1991], Sanov's Theorem is used to show that in such a situation, the probability that the criterion prefers the wrong model ("error probability") decreases exponentially in the sample size. If the Bayesian MAP model selection criterion is used then the following happens: if the data are generated using Poisson[μ] then the error probability decreases exponentially in the sample size, with some error exponent; if the data are generated with Geom[μ] then the overall probability is exponentially decreasing with the same exponent [Cover and Thomas 1991, Theorem 12.9.1 on page 312 and text thereafter]. Thus, we expect that the line for the "known μ" criterion is straight on a logarithmic scale, with a slope that is equal whether the generating distribution is Poisson or geometric. This proves to be the case, as can be seen from Figure 2.

Tests We perform three types of test on the selection criteria, which are described in detail in the following subsections:

1. Error probability measurements.
2. Bias measurements.
3. Calibration testing.

5.1 Error probability

The *error probability* for a criterion is the probability that it will select a model that does not contain the distribution from which the data were sampled. In our experiments, samples are always drawn from a Poisson[μ] distribution with probability p, or from a Geom[μ] distribution with probability $1 - p$. We measure the error probability through repeated sampling; strictly speaking we thus obtain *error frequencies* which approximate the error probability.

Figures 2, 4, 5 and 6 plot the sample size against the error frequency, using different means μ and different priors p on the generating distribution. We use a logscale, which allows for easier comparison of the different criteria; as we pointed out earlier, for the known μ criterion we should expect to obtain a straight line.

In Figures 4, 5 and 6 the log of the error frequency of the known μ criterion is subtracted from the logs of the error frequencies of the other criteria. This brings out the differences in performance in even more detail. The known μ criterion, which has no bias, is perfectly calibrated (as we will observe later) and which also has a low error probability under all circumstances (although biased criteria can sometimes do better if the bias is in the right direction), is thus treated as a baseline of sorts.

5.2 Bias

We define the level of evidence in favour of the Poisson model as:

$$\Delta\left(x^n\right) := L_{\mathrm{G}}(x^n \mid \mu) - L_{\mathrm{P}}(x^n \mid \mu)\,, \tag{22}$$

which is the difference in codelengths according to the known μ criterion. The other criteria define estimators for this quantity: the estimator for a criterion C is defined as:

$$\Delta_{\mathrm{C}}(x^n) := L_{\mathrm{G,C}}(x^n) - L_{\mathrm{P,C}}(x^n) \tag{23}$$

(Some people are more familiar with Bayes factors, of which this is the logarithm.) In our context the *bias* of a particular criterion is the expected difference between the level of evidence according to that criterion and the true level of evidence,

$$E\left[\Delta_{\mathrm{C}}(X^n) - \Delta\left(X^n\right)\right] \tag{24}$$

The value of this expectation depends on the generating distribution, which is assumed to be some mixture of the Poisson and geometric distributions of the same mean.

We measure the bias through sampling. We measure the bias with generating distributions Poisson[8] and Geom[8]; as before we vary the sample size. The results are in Figure 3.

5.3 Calibration

The classical interpretation of probability is frequentist: an event has probability p if in a repeated experiment the frequency of the event converges to p. This interpretation is no longer really possible in a Bayesian framework, since prior assumptions often cannot be tested in a repeated experiment. For this reason, calibration testing is avoided by some Bayesians who may put forward that it is a meaningless procedure from a Bayesian perspective. On the other hand, we take the position that even with a Bayesian hat on, one would like one's inference procedure to be calibrated – in the *idealised* case in which identical experiments are performed repeatedly, probabilities should converge to frequencies. If they do not behave as we would expect even in this idealised situation, then how can we trust inferences based on such probabilities in the real world with all its imperfections?

In the introduction we have indicated the correspondence between codelengths and probability. If the universal codelengths for the different criteria correspond to probabilities that make sense in a frequentist way, then the Bayesian a posteriori probabilities of the two models should too. To test this, we generate samples with a fixed mean and a fixed sample size; half of the samples is drawn from a Poisson distribution and half from a geometric distribution. We then compute the a posteriori probability that it is generated by the Poisson model, for each of the selection criteria. The samples are distributed over 40 bins by discretising their a posteriori probability. For each bin we count the number of sequences that actually were generated by Poisson. If the a posteriori Bayesian probability that the model is Poisson makes any sense in a frequentist way, then the result should be a more or less straight diagonal.

The results are in Figure 7. We used mean 8 and sample size 8 because on the one hand we want a large enough sample size that the posterior has converged to something reasonable, but on the other hand if we choose the sample size even larger it becomes exceedingly unlikely that a sequence is generated of which the probability that it is Poisson is estimated near 0.5, so we would need to generate an infeasibly large number of samples to get accurate results. Note that the "known μ" criterion is perfectly calibrated, because its implicit prior distribution on the mean of the generating distribution puts all probability on the actual mean, so the prior perfectly reflects the truth in this case. Under such circumstances Bayesian and frequentist probability become the same, and we get a perfect answer.

We feel that calibration testing is too often ignored, while it can safeguard against inferences or predictions that bear little relationship to the real world. Moreover, in the objective Bayesian branch of Bayesian statistics, one does emphasise procedures with good frequentist behaviour [Berger 2004]. At least in restricted contexts [Clarke and Barron 1990; Clarke and Barron 1994], Jeffreys' prior has the property that the Kullback-Leibler divergence between the true distribution and the posterior converges to zero quickly, no matter what the true distribution is. Consequently, after observing only a limited number of outcomes, it should already be possible to interpret the posterior as an almost "classical" distribution in the sense that it can be verified by frequentist experiments [Clarke and Barron 1990].

6 Discussion of results

Roughly, the results of our tests can be summarised as follows:

- In this toy problem, as one might expect, all criteria perform extremely well even while the sample size is small. But there are also small but distinct differences that illustrate relative strengths and weaknesses of the different methods. When extrapolated to a more complicated model selection problem, our results should help to decide which criteria are appropriate for the job.
- As was to be be expected, the known μ criterion performs excellently on all tests.
- The criteria based on the plug-in predictive code and BIC/ML exhibit the worst performance.
- The basic restricted ANML criterion yields results that range from good to very bad, depending on the chosen parameter range. Since the range must be chosen without any additional knowledge of the properties of the data, this criterion is a bit arbitrary.
- The results for the two-part restricted ANML and Objective Bayesian criteria are reasonable in all tests we performed; these criteria thus display robustness.

In the following subsections we evaluate the results for each model selection criterion in turn.

6.1 Poor performance of the plug-in criterion

One feature of Figure 2 that immediately attracts attention is the unusual slope of the error rate line of the plug-in criterion, which clearly favours the geometric distribution. This is even clearer in Figure 3, where the plug-in criterion can be observed to become more and more favourable to the geometric model as the sample size increases, regardless of whether the data were sampled from a Poisson or geometric distribution. This is also corroborated by the results on the calibration test, where the plug-in criterion most severely underestimates the probability that the data was sampled from a Poisson distribution: of the sequences assessed to be Poisson with only 20% probability, in fact about 60% turned out to be sampled from a Poisson distribution.

While this behaviour may seem appropriate if the data are judged more likely to come from a geometric distribution, there is actually a strong argument that *even under those circumstances it is not the most desirable behaviour*, for the following reason. Suppose that we put a fixed prior on the generating distribution, with nonzero probability for both distributions. The marginal error probability is a linear combination of the probabilities of error for the two generating distributions; as such it is dominated by the probability of error with the *worst* exponent. So if minimising the error probability is our goal, then we must conclude that the behaviour of the plug-in criterion is suboptimal. (On a side note, minimising the error probability with respect to a fixed prior is *not* the goal of classical hypothesis testing, since in that setting the two hypotheses do not play a symmetrical role.) To illustrate, the bottom graph in Figure 4 shows that, even if there is only a 10% chance that the data are Poisson, then the plug-in criterion still has a worse (marginal) probability of error than "known μ" as soon as the sample size reaches 25. Figure 5 shows what happens if the prior on the generating distribution is uniform – using the plug-in criterion immediately yields the largest error probability of all the criteria under consideration. This effect only becomes stronger if the mean is higher.

This strangely poor behaviour of the plug-in criterion initially came as a complete surprise to us. Theoretical literature certainly had not suggested it. Rather the contrary: in [Rissanen 1989] we find that "it is only because of a certain inherent singularity in the process [of plug-in coding], as well as the somewhat restrictive requirement that the data must be ordered, that we do not consider the resulting predictive codelength to provide another competing definition for the stochastic complexity, but rather regard it as an approximation". There are also many results showing that the regret for the plug-in code grows as $\frac{k}{2}\ln n$, the same as the regret for the NML code, for a variety of models. Examples are [Rissanen 1986; Gerensce'r 1987; Wei 1990]. Finally, publications such as [Modha and Masry 1998; Kontkanen et al. 2001] show excellent behaviour of the plug-in criterion for model selection in regression and classification based on Bayesian networks, respectively. So, we were extremely puzzled by these results at first.

To gain intuition as to why the plug-in code should behave so strangely, note that the variance of a geometric distribution is much larger than the variance of the Poisson distribution with the same mean. This suggests that the penalty for using $\hat{\mu}$ rather than μ to code each consecutive outcome is higher for the Poisson model. The accumulated difference accounts for the difference in regret.

We have made this intuition precise in a separate publication. We prove in [Grünwald and de Rooij 2005] that for single parameter exponential families, the regret for the plug-in code grows with $\frac{1}{2}\ln(n)\mathrm{Var}_P(X)/\mathrm{Var}_M(X)$, where n is the sample size, P is the generating distribution and M is the best element of the model (the element of $\mathcal{M}$ for which the Kullback Leibler divergence $D(P \parallel M)$ is minimised). The plug-in model has the same regret (to $O(1)$) as the NML model if and only if the variance of the generating distribution is the same as

the variance of the best element of the model. The existing literature studies the case where $M = P$, so automatically $\mathrm{Var}_M(X) = \mathrm{Var}_P(X)$.

We should add here that theoretical arguments [Barron et al. 1998] show that there are quite strong limits to how badly the plug-in criterion can behave. For example, whenever a finite or countably infinite set of parametric models (each containing an uncountable number of distributions) are being compared, and data are i.i.d. according to an element of one of the models, then the error probability of the plug-in criterion *must* go to 0. If the number of models is finite and they are non-nested, it must even go to 0 as $\exp(-cn)$ for some constant $c > 0$. The same holds for other criteria including BIC, but not necessarily for ML. The plug-in criterion may have slightly lower c than other model selection procedures, but the ML criterion is guaranteed to fail (always select the most complex model) in cases such as regression with polynomials of varying degree, where the number of models being compared is nested and countably infinite. Thus, whereas in our setting the plug-in criterion performs somewhat worse (in the sense that more data are needed before the same quality of results is achieved) than the ML criterion, it is guaranteed to display reasonable behaviour over a wide variety of settings, in many of which the ML criterion fails utterly.

Similarly, we suspect that in settings involving a countably infinite number of models, the plug-in criterion will often also outperform the BIC criterion (see below), which is based on an approximation of the Bayesian posterior that can be arbitrarily bad. Empirical evidence for this in the context of Bayesian network model selection for classification is provided by [Kontkanen et al. 2001] and [Allen, Madani, and Greiner 2003].

In some cases, the plug-in criterion may even be preferable to Bayesian model selection, since for many model families, the Bayesian integral (11) cannot be evaluated analytically. It is then often approximated by, for example, Markov Chain Monte Carlo methods, and it is not at all clear whether the resulting procedure will show better or worse performance than the plug-in criterion.

Thus, we would like to emphasise that we do *not* discourage the use of the plug-in criterion. It has some good theoretical properties which show that in idealised situations, it must eventually select the 'correct' model. However, our results do indicate that the plug-in criterion should be used with caution, and may exhibit worse performance than selection criteria that are more robust under misspecification.

6.2 ML/BIC

Beside known μ and plug-in, all criteria seem to share more or less the same error exponent. Nevertheless, they still show differences in bias. While we have to be careful to avoid over-interpreting our results, we find that the ML/BIC criterion consistently displays the largest bias in favour of the Poisson model. Figure 3 shows how the Poisson model is *always* at least $10^{0.7} \approx 5$ times more likely according to ML/BIC than according to known μ, regardless whether data were sampled from a geometric or a Poisson distribution. Figure 7 contains further evidence of bias in favour of the Poisson model: together with the plug-in criterion, the ML/BIC criterion exhibited the worst calibration performance: when the probability that the data is Poisson distributed is assessed by the ML criterion to be around 0.5, the real frequency of the data being Poisson distributed is only about 0.2.

This illustrates how the Poisson model appears to have a greater descriptive power, even though the two models have the same number of parameters, an observation which we hinted at in Section 3. Intuitively, the Poisson model allows more information about the data to be stored in the parameter estimate. All the other selection criteria compensate for this effect, by giving a higher probability to the geometric model. (In terms of coding, the Poisson codelength

is increased by more than the geometric codelength.)

6.3 Basic restricted ANML

We have seen that the ML/BIC criterion shows the largest bias for the Poisson model. Figure 3 shows that the second largest bias is achieved by ANML $\mu^* = 10$. Apparently the correction term that is applied by ANML criterion is not sufficient if we choose $\mu^* = 10$. However, we can obtain any correction term we like since we observed in Section 4.2 that ANML is equivalent to a GLRT with a selection threshold that is an unbounded, monotonically increasing function of μ^*. Essentially, by choosing an appropriate μ^* we can get *any* correction in favour of the geometric model, even one that would lead to a very large bias in the direction of the geometric model. We conclude that it does not really make sense to use a fixed restricted parameter domain to repair the NML model when it does not exist, unless prior knowledge is available.

6.4 Objective Bayes and two-part restricted ANML

We will not try to interpret the differences in error probability for the (approximated) Bayesian and ANML 2-part criteria. Since we are using different selection criteria we should expect at least some differences in the results. These differences are exaggerated by our setup with its low mean and small sample size.

The Bayesian criterion, as well as its approximation appear to be somewhat better calibrated than the two-part ANML but the evidence is too thin to draw any strong conclusions.

Figures 4–6 show that the error probability for these criteria tends to decrease at a slightly lower rate than for known μ (except when the prior on the generating distribution is heavily in favour of Poisson). While we do not understand this phenomenon well enough so as to prove it mathematically, it is of course consistent with the general rule that with more prior uncertainty, more data are needed to make the right decision. It may be that all the information contained within a sample can be used to improve the resolution of the known μ criterion, while for the other criteria some of that information has to be sacrificed in order to estimate the parameter value.

7 Summary and conclusion

We have performed error probability tests, bias tests and calibration tests to study the properties of a number of model selection criteria. These criteria are based on the MDL philosophy and involve computing the codelength of the data with the help of the model. There are several ways to compute such a code, but the preferred method, the Normalised Maximum Likelihood (NML) code, cannot be applied since it does not exist for the Poisson and geometric models that we consider.

We have experimented with the following alternative ways of working around this problem: (1) using BIC which is a simplification of approximated NML (ANML), (2) ANML with a restricted parameter range, this range being either fixed or encoded separately, (3) a Bayesian model using Jeffreys' prior, which is improper for the case at hand but which can be made proper by conditioning on the first outcome of the sample, (4) its approximation and (5) a plug-in code which always codes the new outcome using the distribution indexed by the maximum likelihood estimator for the preceding outcomes.

Both BIC and ANML with a fixed restricted parameter range define a GLRT test and can be interpreted as methods to choose an appropriate threshold. BIC implies a neutral threshold, so the criterion will become biased in favour of the model which is most susceptible to overfitting.

We found that even though both models under consideration have only one parameter, a GLRT with neutral threshold tends to be biased in favour of Poisson. ANML implies a threshold that counteracts this bias, but for every such threshold value there exists a corresponding parameter range, so it does not provide any more specific guidance in selecting that threshold. If the parameter range is separately encoded, this problem is avoided and the resulting criterion behaves competitively, although it is not calibrated as well as the Bayesian criterion and the two-part codelength is slightly redundant.

The Bayesian criterion displays reasonable performance both on the error rate experiments and the calibration test. The Bayesian universal codes for the models are not redundant and admit an MDL interpretation as minimising worst-case codelength in an expected sense (Section 4.7).

The plug-in criterion has a bias in favour of the geometric model that depends strongly on the sample size. As a consequence its error rate decreases more slowly in the sample size if we put a prior on the generating distribution that assigns nonzero probability to both models. This result was surprising to us and has led to a theoretical analysis of the codelength of the plug-in code in [Grünwald and de Rooij 2005]. It turns out that the regret of the plug-in code does not necessarily grow with $\frac{k}{2} \ln n$ like the NML and Bayesian codes do, if the sample is not distributed according to any element of the model. We conjecture that model selection based on plug-in codes continues to behave suboptimally in more general settings. However, we should note that there are strong limits to 'how bad things can get'. Various results [Barron et al. 1998] indicate that model selection based on plug-in codes must eventually select the correct model (if such a model exists), even when the number of models under consideration is unbounded.

In conclusion, while NML certainly seems a sensible approach to defining model selection criteria, when it is undefined it is impossible to minimise the worst case regret. There are many different methods to deal with this problem, some of which work reasonably well and some of which work surprisingly badly. The fundamental question remains: if it is not possible to minimise the worst case regret, then what exactly *should* we optimise?

Acknowledgements

The main idea for this article is not our own, but comes from Aaron D. Lanterman's text "Hypothesis Testing for Poisson versus Geometric Distributions using Stochastic Complexity" [Lanterman 2005] which it is a pleasure to read. He deserves much credit. Furthermore we wish to thank the editors: their extensive, insightful and fair criticisms have led to significant improvements in the text.

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.

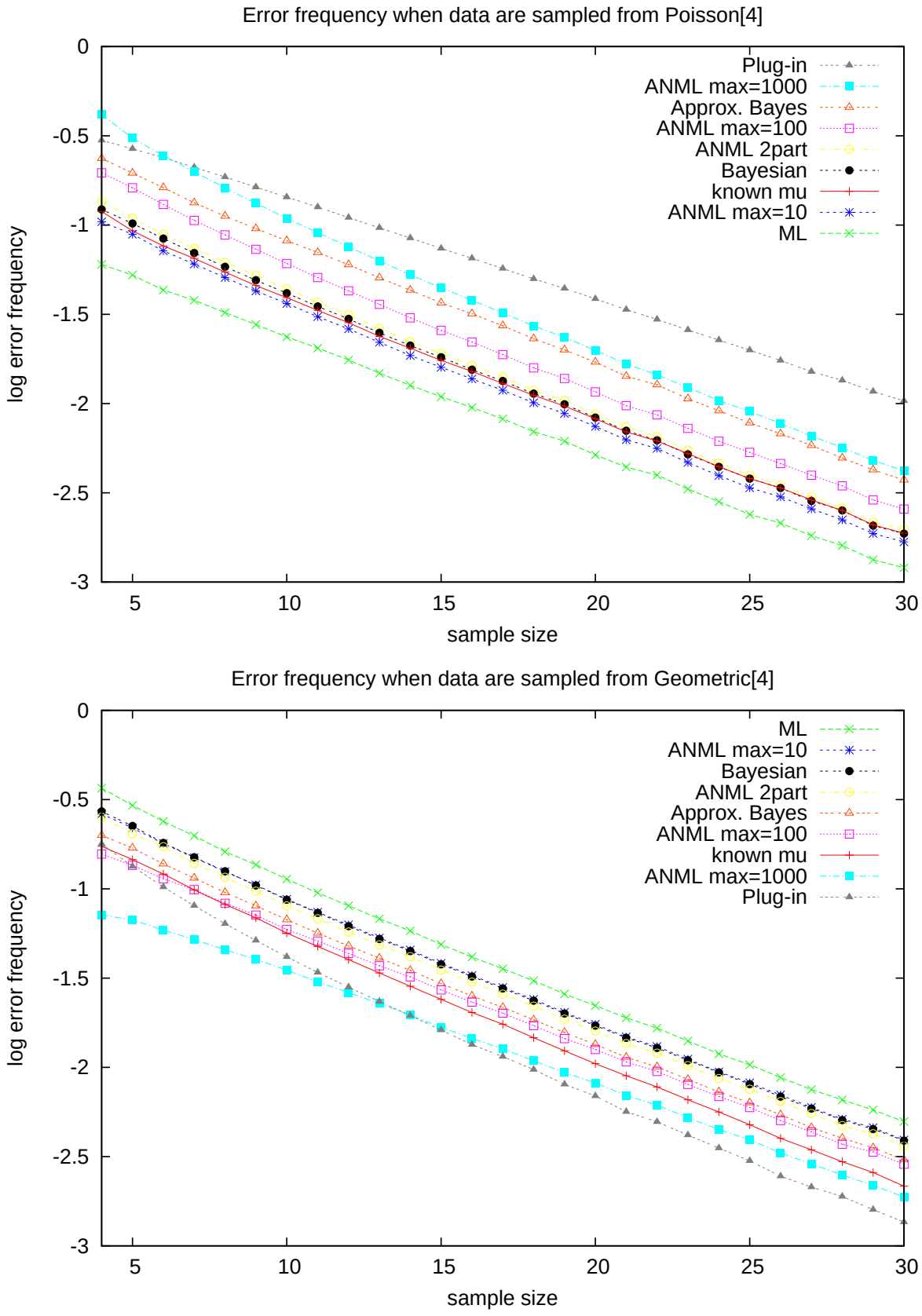

Figure 2: The $\log_{10}$ of the error frequency.

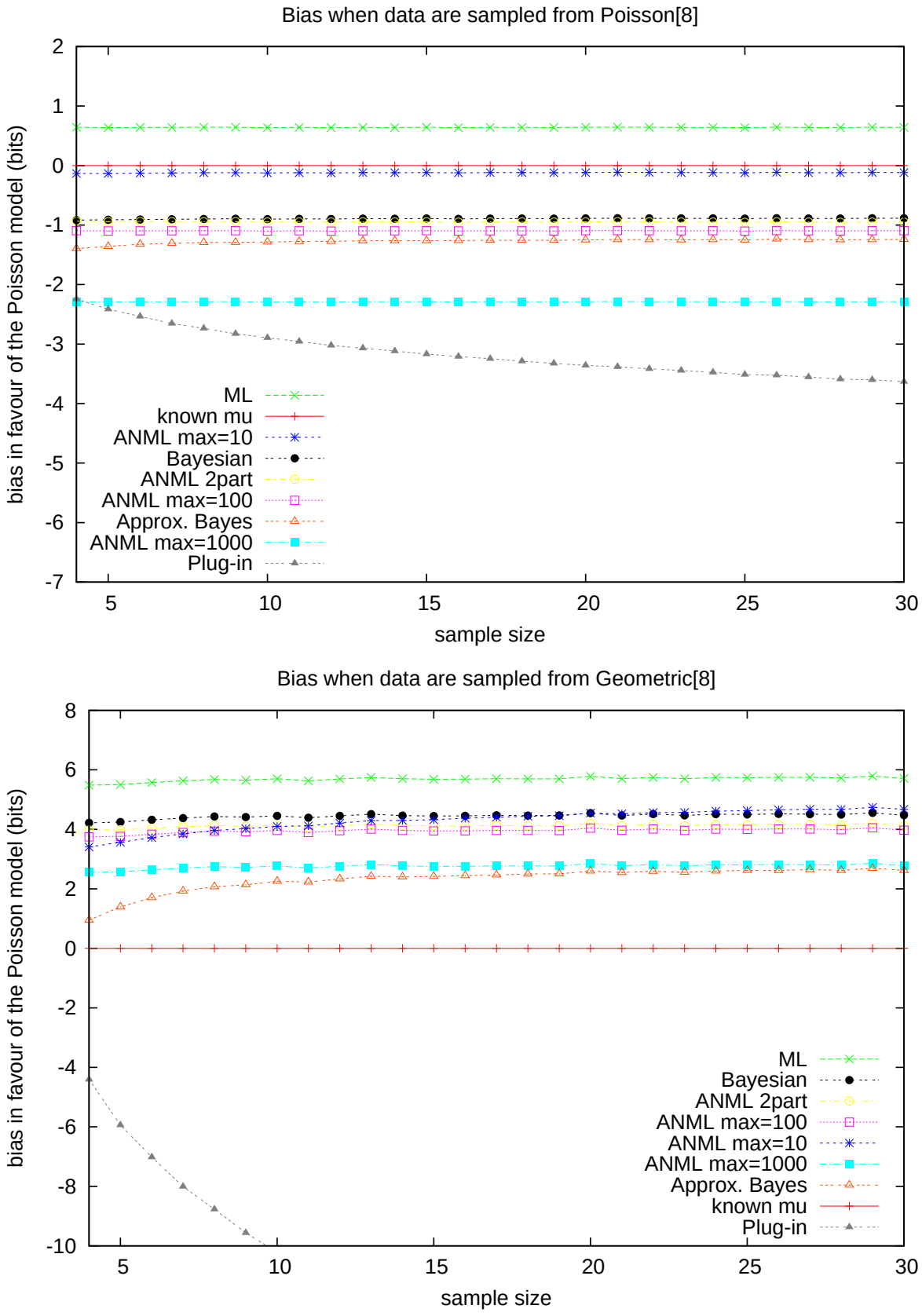

Figure 3: The classification bias in favour of the Poisson model in bits.

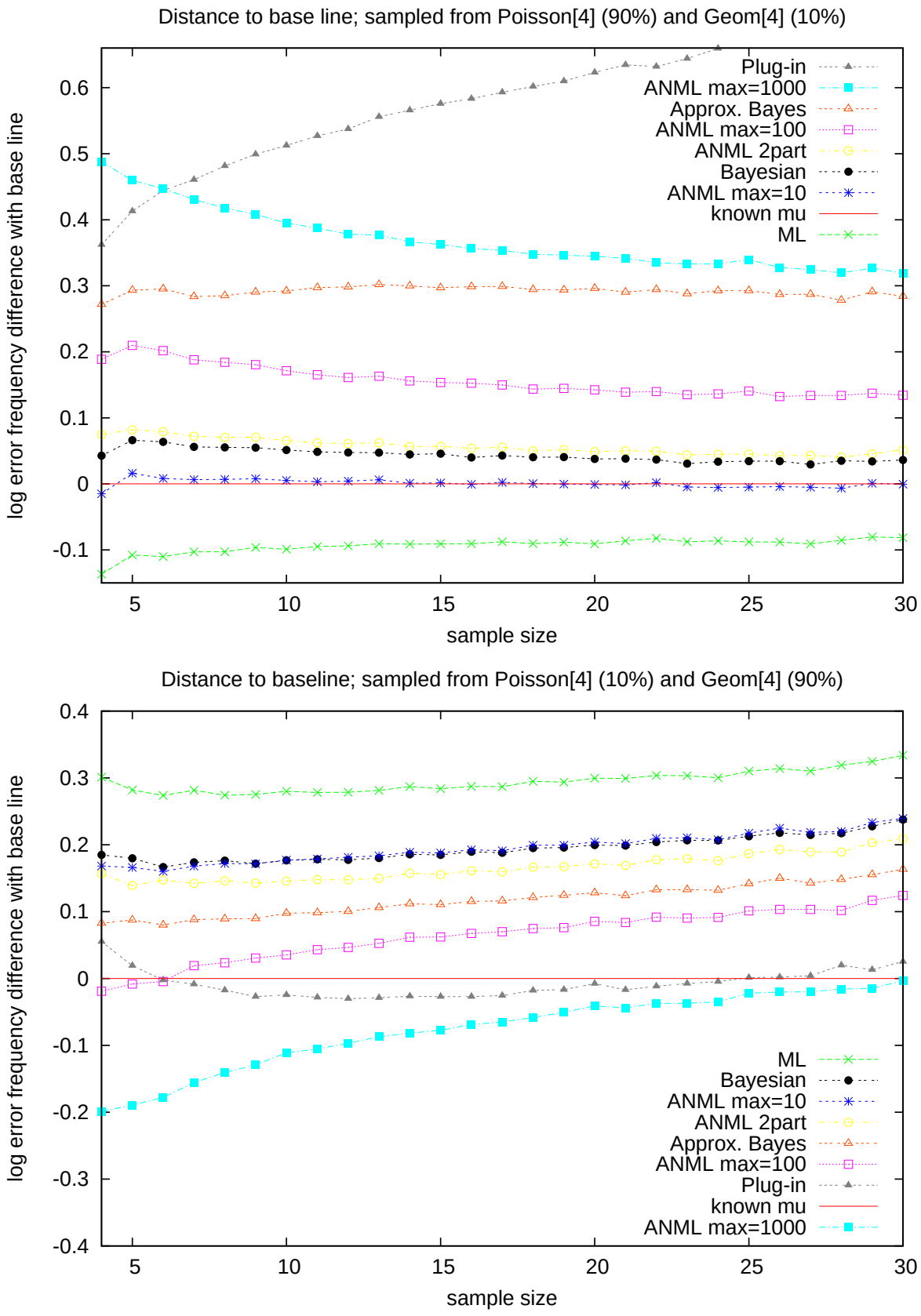

Figure 4: The difference in the $\log_{10}$ of the frequency of error between each criterion and the known μ criterion. The mean is 4. In these graphs, data are sampled from one of the two models with unequal probabilities.

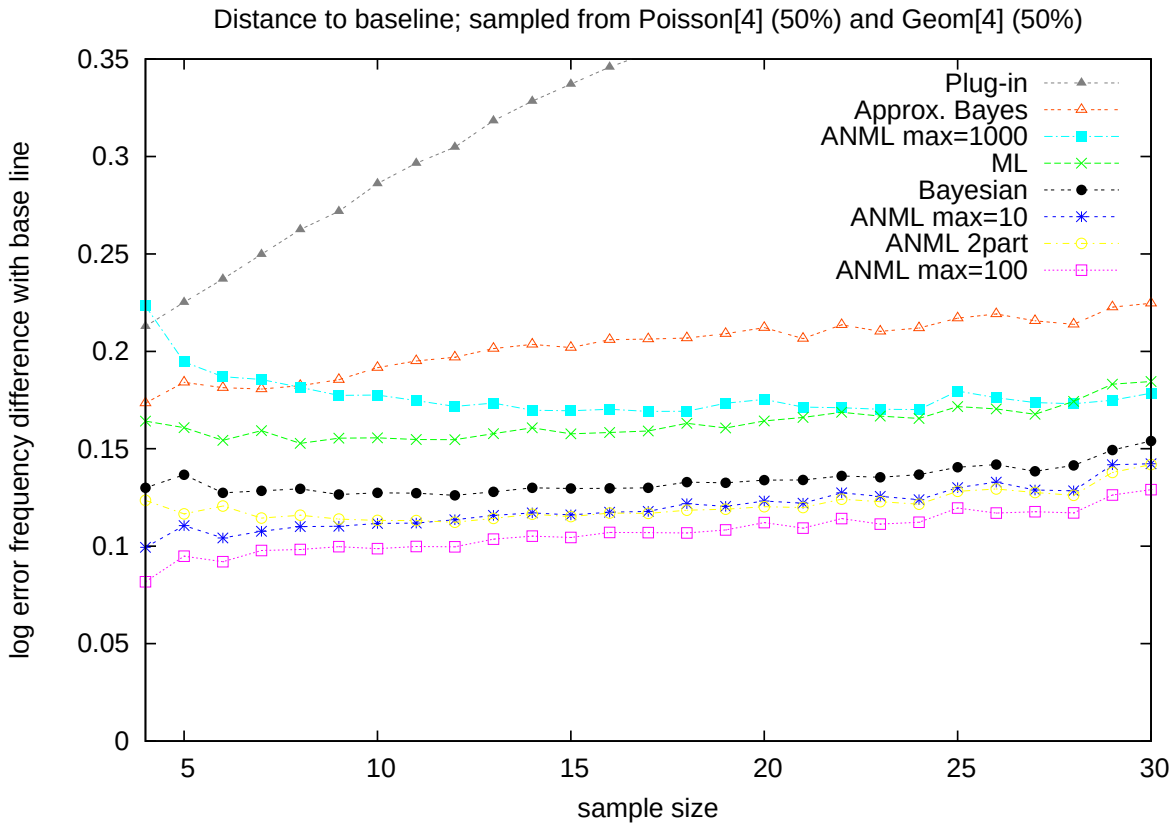

Figure 5: The difference in the $\log_{10}$ of the frequency of error between each criterion and the known μ criterion. The mean is 4.

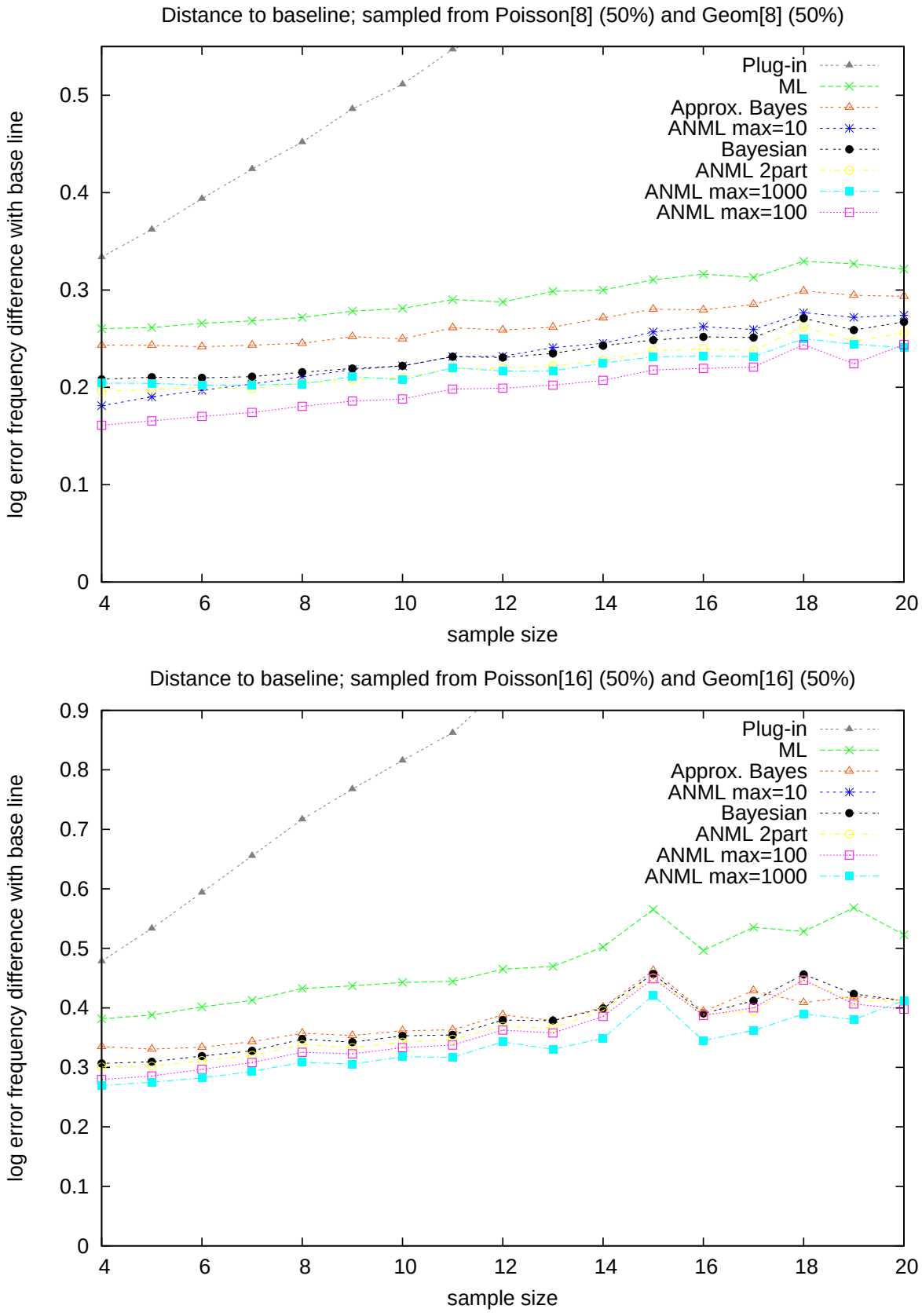

Figure 6: The difference in the $\log_{10}$ of the frequency of error between each criterion and the known μ criterion. The mean is 8 in the top graph and 16 in the bottom graph.

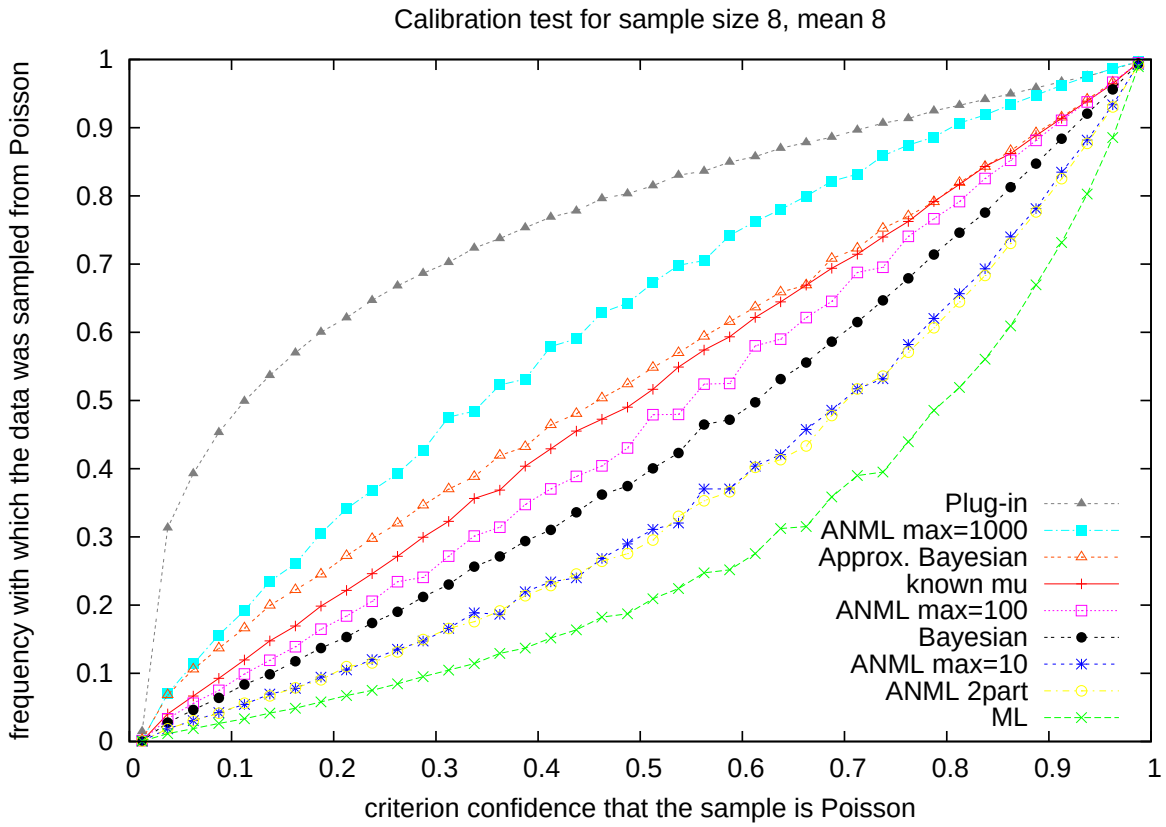

Figure 7: Calibration: probability that the model assigned to the data being Poisson against the frequency with which it actually was Poisson.

References

Allen, T. V., O. Madani, and R. Greiner (2003, July). Comparing model selection criteria for belief networks. submitted.

Balasubramanian, V. (1997). Statistical inference, Occam's Razor, and statistical mechanics on the space of probability distributions. *Neural Computation 9*, 349–368.

Barron, A., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory 44*(6), 2743–2760. Special Commemorative Issue: Information Theory: 1948-1998.

Berger, J. (2004). Personal communication.

Berger, J. O. and L. R. Pericchi (1997). Objective Bayesian methods for model selection: introduction and comparison. *Institute of Mathematical Statistics Lecture Notes (Monograph series) 38*, 135–207.

Bernardo, J. and A. F. Smith (1994). *Bayesian Theory*. Wiley.

Chater, N. (2005). A Minimum Description Length principle for perception. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Clarke, B. and A. Barron (1990). Information theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory 36*, 453–471.

Clarke, B. and A. Barron (1994). Jeffreys' prior is asymptotically least favourable under entropy risk. *The Journal of Statistical Planning and Inference 41*, 37–60.

Cover, T. and J. Thomas (1991). *Elements of Information Theory*. New York: Wiley Interscience.

Dawid, A. (1984). Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A 147*(2), 278–292.

Gerensce'r, L. (1987). Order estimation of stationary Gaussian ARMA processes using Rissanen's complexity. Technical report, Computer and Automation Institute of the Hungarian Academy of Sciences.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology* (44), 133–152.

Grünwald, P. and S. de Rooij (2005). Asymptotic log-loss of prequential maximum likelihood codes. In *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT 2005)*, pp. 652–667. Springer.

Grünwald, P. D. (2005). MDL tutorial. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Grünwald, P. D., I. J. Myung, and M. A. Pitt (Eds.) (2005). *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Jeffreys, H. (1961). *Theory of Probability* (Third ed.). London: Oxford University Press.

Kontkanen, P., P. Myllymäki, and H. Tirri (2001). Comparing prequential model selection criteria in supervised learning of mixture models. In T. Jaakkola and T. Richardson (Eds.), *Proceedings of the Eighth International Conference on Articial Intelligence and Statistics*, pp. 233–238. Morgan Kaufman.

Lanterman, A. D. (2005). Hypothesis testing for Poisson versus geometric distributions using stochastic complexity. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Lee, M. and D. Navarro (2005). Minimum description length and psychological clustering models. In P. Grünwald, I. Myung, and M. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Liang, F. and A. Barron (2005). Exact minimax predictive density estimation and MDL. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. MIT Press.

Modha, D. S. and E. Masry (1998). Prequential and cross-validated regression estimation. *Machine Learning 33*(1).

Myung, I., V. Balasubramanian, and M. Pitt (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA 97*, 11170–11175.

Myung, I. J., M. A. Pitt, S. Zhang, and V. Balasubramanian (2001). The use of MDL to select among computational models of cognition. In *Advances in Neural Information Processing Systems*, Volume 13, pp. 38–44. MIT Press.

Navarro, D. (2004). A note on the applied use of mdl appproximations. *Neural Computation 16*.

Pitt, M., I. Myung, and S. Zhang (2002). Toward a method of selecting among computational models of cognition. *Psychological Review* (3), 472–491.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica 14*, 465–471.

Rissanen, J. (1984). Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory 30*, 629–636.

Rissanen, J. (1986). A predictive least squares principle. *IMA Journal of Mathematical Control and Information 3*, 211–222.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company.

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

Rissanen, J. (2000). MDL denoising. *IEEE Transactions on Information Theory 46*(7), 2537–2543.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Wagenmakers, E., P. Grünwald, and M. Steyvers (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*. (this issue).

Wei, C. (1990). On predictive least squares principles. *Annals of Statistics 20*(1), 1–42.