Minimum Description Length Model Selection

de Rooij, S.

Publication date
2008

Chapter 2

Dealing with Infinite Parametric Complexity

Model selection is the task of choosing one out of a set of hypotheses for some phenomenon, based on the available data, where each hypothesis is represented by a model, or set of probability distributions. As explained in the introductory chapter, MDL model selection proceeds by associating a so-called *universal distribution* with each model. The number of bits needed to encode the data D using the universal code associated with model $\mathcal{M}$ is denoted $L_{\mathcal{M}}(D)$. We then pick the model that minimises this expression and thus achieves the best compression of the data.

Section 1.2.3 lists the most important such universal codes, all of which result in slightly different code lengths 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.

The MDL literature emphasises that the used universal code should be efficient *whatever data are observed*: whenever the model contains a code or distribution that fits the data well in the sense that it achieves a small code length, the universal code length should not be much larger. More precisely put, the universal code should achieve small regret in the worst case over all possible observations, where the regret is the difference between the universal code length and the shortest code length achieved by an element of the model $\mathcal{M}$. The *normalised maximum likelihood* (NML) code (Section 1.2.3) minimises this worst-case regret. Moreover, the NML regret is a function of only the sample size n, and does not depend on the data. As such it can be viewed as an objective measure of model complexity: the *parametric complexity* of a model is the regret achieved by the NML code, as a function of the sample size n. For this reason NML is the preferred universal code in modern MDL.

However, it turns out that the parametric complexity is infinite for many models, so that NML-based model selection is not effective. Other universal codes

that do achieve finite code lengths can usually still be defined for such models, but those codes have a worst-case regret that depends not only on the sample size n, but also on the parameter of the best fitting distribution in the model $\hat{\theta}$; such codes are *luckiness codes* in the sense of Section 1.1.1. Model selection criteria based on such luckiness codes thus acquire an element of subjectivity.

Several remedies have been proposed for this situation: variations on NML-based model selection that can obviously not be worst case optimal, but that do achieve finite code length without explicitly introducing regret that depends on the element of the model that best fits the data. In this chapter we evaluate such alternative procedures empirically for model selection between the simple Poisson and geometric models. We chose these two models since they are just about the simplest and easiest-to-analyse 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 prequential maximum likelihood code – which was found to perform quite well in some contexts [61, 52] – exhibits poor behaviour in our experiments. We briefly discuss the reasons for this behaviour in Section 2.5; a full analysis is the subject of Chapter 3.

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 2.3.6, 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 2.1 we describe the code that achieves worst-case minimal regret. This code does not exist for the Poisson and geometric distributions. We analyse these models in more detail in Section 2.2.

In Section 2.3 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 2.4. The results are evaluated in Section 2.5. Our conclusions are summarised in Section 2.6.

2.1 Universal codes and Parametric Complexity

We first briefly review some material from the introductory chapter that is especially important to this chapter. The *regret* of a code L with respect to a parametric model $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ with parameter space Θ on a sequence of outcomes $x^n = x_1, \ldots, x_n \in \mathcal{X}^n$ is

$$\mathcal{R}(L, \mathcal{M}, x^n) := L(x^n) - \inf_{\theta \in \Theta} - \log P_\theta(x^n),$$

which was first defined in (1.4), page 5. A function $\hat{\theta} : \mathcal{X}^n \to \Theta$ that maps outcomes x^n to a parameter value that achieves this infimum is called a *maximum*

likelihood estimator. We sometimes abbreviate $\hat{\theta} = \hat{\theta}(x^n)$.

A code L is *f-universal with respect to model $\mathcal{M}$* if $\mathcal{R}(L, \mathcal{M}, x^n) \leq f(\hat{\theta}, n)$ for all n and all x^n (Definition 1.2.2, page 14). The MDL philosophy [72, 38] has it that the best universal code minimises the regret 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 in (1.13), page 18:

$$P_{\text{nml}}(x^n) := \frac{P(x^n|\hat{\theta}(x^n))}{\sum_{y^n} P(y^n|\hat{\theta}(y^n))},$$

with corresponding code length

$$L_{\text{nml}}(x^n) := -\log P(x^n|\hat{\theta}(x^n)) + \log \sum_{y^n} P(y^n|\hat{\theta}(y^n)).$$

This code length is called the *stochastic complexity* of the data. The first term in the equation is the code length of the data using the maximum likelihood estimator; the second term is the $-\log$ of the normalisation factor of P_{nml}, called the *parametric complexity* of the model $\mathcal{M}$ at sample size n.

It is usually impossible to compute the parametric complexity analytically, but there exists a good approximation L_{anml}, due to Rissanen, Takeuchi and Barron [72, 88, 89, 87]:

$$L_{\text{anml}}(x^n) := L(x^n|\hat{\theta}) + \frac{k}{2}\log\frac{n}{2\pi} + \log\int_{\Theta}\sqrt{\det I(\theta)}\,\mathrm{d}\theta. \tag{2.1}$$

Here, n, k and $I(\theta)$ denote the number of outcomes, the number of parameters and the Fisher information matrix respectively. If $\mathcal{M}$ is an exponential family model such as the Poisson and Geometric models considered in this chapter, and is parameterised by $\Theta \subseteq \mathbb{R}^k$ for some $k \in \mathbb{Z}^+$, and if both the parametric complexity and the integral in (2.1) are finite, then we have the following. For any Θ' with nonempty interior, and whose closure is a bounded subset of the interior of Θ, we have

$$\lim_{n\to\infty} \sup_{x^n:\hat{\theta}(x^n)\in\Theta'} |L_{\text{nml}}(x^n) - L_{\text{anml}}(x^n)| = 0.$$

Thus, the approximation uniformly converges to the exact code length for all sequences of outcomes whose maximum likelihood estimator does not converge to the boundaries of the parameter space. For more information, see [39]. Since the last term in (2.1) 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, (2.1) 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.

As stated in the introduction, the problem is that for many models the parametric complexity is infinite. Many strategies have been proposed to deal with this, but most are somewhat ad-hoc. When Rissanen defines stochastic complexity as $L_{\mathrm{nml}}(x^n)$ in [72], he writes that he does so "thereby concluding a decade long search", but as Lanterman observes in [53], "in the light of these problems we may have to postpone concluding the search just a while longer".

2.2 The Poisson and Geometric Models

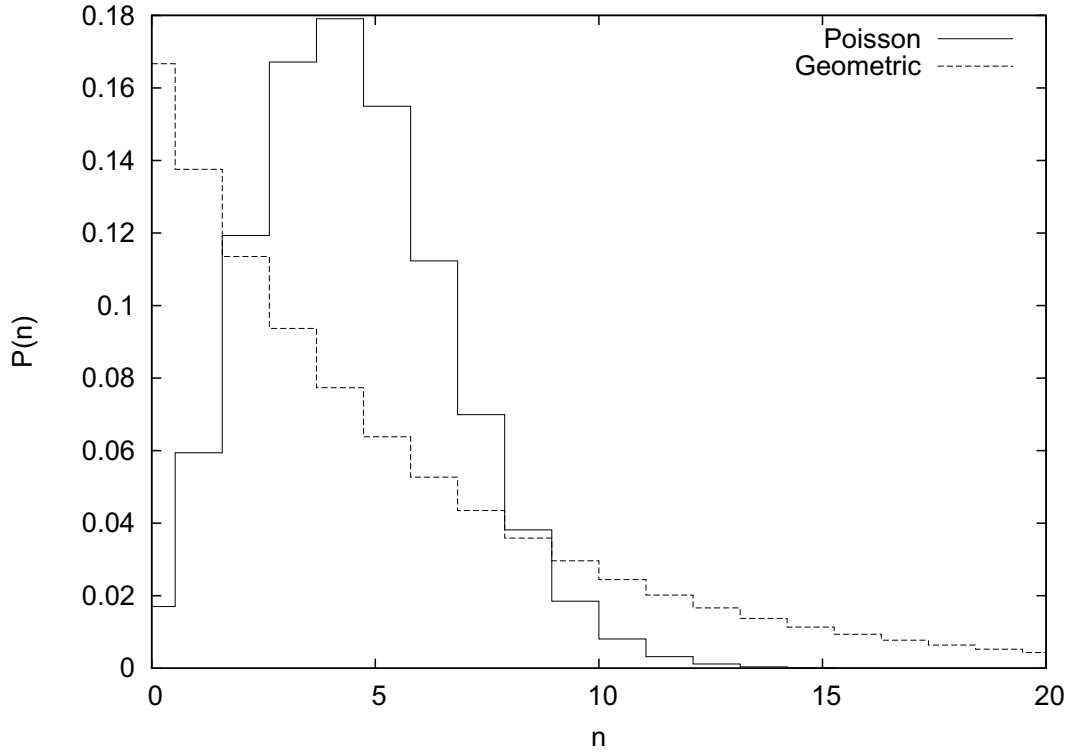
We investigate MDL model selection between the Poisson and Geometric models. Figure 2.1 may help form an intuition about the probability mass functions of the two distributions. One reason for our choice of models is that they are both single parameter models, so that the dominant $\frac{k}{2}\log\frac{n}{2\pi}$ term of (2.1) cancels. This means that at least for large sample sizes, simply picking the model that best fits the data should always work. We nevertheless observe that for small sample sizes, data generated by the geometric distribution are misclassified as Poisson much more frequently than the other way around (see Section 2.5). 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 favour of Poisson that is introduced if we just select the best fitting model, we would like to compute the third term of (2.1), 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 2.3 we describe a number of methods that have been proposed in the literature as ways to deal with infinite parametric complexity; in Section 2.5 they are evaluated empirically.

Reassuringly, all methods we investigate tend to compensate for this overfitting phenomenon by "punishing" the Poisson model. However, to what extent the bias is compensated depends on the used method, 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, parameterisation by the mean is standard. For geometric, the reparameterisation can be arrived at by noting that in the standard parameterisation, $P(x|\theta) = (1-\theta)^x\theta$, the mean is given by $\mu = (1-\theta)/\theta$. As a notational reminder the parameter is called μ henceforth. Conveniently, the ML estimator

Figure 2.1 The mean 5 Poisson and geometric distributions.



$\hat{\mu}$ for both distributions is the average of the data.

We will add a subscript P or G to indicate that code lengths are computed with respect to the Poisson model or the Geometric model, respectively. Furthermore, to simplify the equations in the remainder of this chapter somewhat we will express code lengths in nats ($-\ln$ probabilities) rather than bits ($-\log_2$ probabilities).

$$L_{\mathrm{P}}\left(x^n|\mu\right) = -\ln\prod_{i=1}^{n}\frac{e^{-\mu}\mu^{x_i}}{x_i!} = \sum_{i=1}^{n}\ln(x_i!) + n\mu - \ln\mu\sum_{i=1}^{n}x_i, \tag{2.2}$$

$$L_{\mathrm{G}}\left(x^n|\mu\right) = -\ln\prod_{i=1}^{n}\frac{\mu^{x_i}}{(\mu+1)^{x_i+1}} = n\ln(\mu+1) - \ln\left(\frac{\mu}{\mu+1}\right)\sum_{i=1}^{n}x_i. \tag{2.3}$$

2.3 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 methods leads to one, or sometimes more, concrete model selection criteria which we evaluate in Section 2.5.

2.3.1 BIC/ML

One way to deal with the diverging integral in 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 as an approximation to the Bayesian marginal likelihood [47], but it became well-known only when it was proposed by Rissanen [67] and Schwarz [78]. 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 code length. We note that Rissanen already abandoned the idea in the mid 1980's in favour of more sophisticated code length approximations. Because of its connection to the Bayesian marginal likelihood, it is best known as the BIC (Bayesian Information Criterion):

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

Comparing BIC to the approximated NML code length 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

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

We are left with a generalised likelihood ratio test (GLRT). In such a test, the ratio of the probabilities under the two models, $P_{\mathrm{P}}\left(x^n|\hat{\mu}\right)/P_{\mathrm{G}}\left(x^n|\hat{\mu}\right)$ 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. As we remarked before, experience shows that this approach often leads to overfitting and 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.)

2.3.2 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 (2.1) stays finite. Our approach is to impose a maximum on the allowed mean by setting $\Theta = (0, \mu^*)$.

To compute the approximated parametric complexity of the restricted models we need to establish the Fisher information first. We use $I(\theta) = -E_\theta[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} L\left(x|\theta\right)]$ to obtain

$$I_{\mathrm{P}}(\mu) = -E_\mu\left[-\frac{x}{\mu^2}\right] = \frac{1}{\mu}, \text{ and} \tag{2.4}$$

$$I_{\mathrm{G}}(\mu) = -E_\mu\left[-\frac{x}{\mu^2} + \frac{x+1}{(\mu+1)^2}\right] = \frac{1}{\mu(\mu+1)}. \tag{2.5}$$

Now we can compute the last term in the parametric complexity approximation (2.1):

$$\ln \int_0^{\mu^*} \sqrt{I_{\mathrm{P}}(\mu)}\,\mathrm{d}\mu = \ln \int_0^{\mu^*} \mu^{-\frac{1}{2}}\,\mathrm{d}\mu = \ln\left(2\sqrt{\mu^*}\right); \tag{2.6}$$

$$\ln \int_0^{\mu^*} \sqrt{I_{\mathrm{G}}(\mu)}\,\mathrm{d}\mu = \ln \int_0^{\mu^*} \frac{1}{\sqrt{\mu(\mu+1)}}\,\mathrm{d}\mu = \ln\left\{2\ln\left(\sqrt{\mu^*} + \sqrt{\mu^*+1}\right)\right\}. \tag{2.7}$$

Thus, the parametric complexities of the restricted models are both monotonically increasing functions of μ^*. Let the function $\delta(\mu^*) := \ln(2\sqrt{\mu^*}) - \ln(2\ln(\sqrt{\mu^*} + \sqrt{\mu^*+1}))$ measure the difference between the parametric complexities. We obtain a model selection criterion that selects the Geometric model if

$$0 < L_{\mathrm{P,ANML}(\mu^*)}(x^n) - L_{\mathrm{G,ANML}(\mu^*)}(x^n) = L_{\mathrm{P}}\left(x^n|\hat{\mu}\right) - L_{\mathrm{G}}\left(x^n|\hat{\mu}\right) + \delta(\mu^*).$$

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

It is not hard to show that the parametric complexity of the restricted Poisson model grows *faster* with μ^* than the parametric complexity of the Geometric model: $\delta(\mu^*)$ is monotonically increasing in μ^*, and grows unboundedly in μ^*. This indicates that the Poisson model has more descriptive power, even though the models have the same number of parameters and both have infinite parametric complexity.

An obvious conceptual problem with restricted ANML is that the imposed restriction is quite arbitrary and requires a priori knowledge about the scale of the observations. But the parameter range can be interpreted as a hyper-parameter, which can be incorporated into the code using several techniques; two such techniques are discussed next.

2.3.3 Two-part ANML

Perhaps the easiest way to get rid of the μ^* parameter that determines the parameter range, and thereby the restricted ANML code length, is to use a two-part

code. The first part contains a specification of μ^*, which is followed by an encoding of the rest of the data with an approximate code length given by restricted ANML for that μ^*. 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). 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 code length of the data itself, but we admit that the code for the parameter range allows much more 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 [38], 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.

2.3.4 Renormalised Maximum Likelihood

Related to the two-part restricted ANML, but more elegant, is Rissanen's *renormalised maximum likelihood* (RNML) code, [73, 38]. 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 complex-

ities are still infinite; we have not performed a second renormalisation. Therefore, the RNML code is not included in our experiments.

2.3.5 Prequential Maximum Likelihood

The *prequential maximum likelihood code*, which we will abbreviate PML-code, is an attractive universal code because it is usually a lot easier to implement than either NML or a Bayesian code. Moreover, its implementation hardly requires any arbitrary decisions. Here the outcomes are coded sequentially using the probability distribution indexed by the ML estimator for the previous outcomes [26, 69]; for a general introduction see [96] or [38].

$$L_{\mathrm{PIPC}}\left(x^n\right) = \sum_{i=1}^{n} L\left(x_i|\hat{\mu}(x^{i-1})\right),$$

where $L\left(x_i|\hat{\mu}(x^{i-1})\right) = -\ln P\left(x_i|\hat{\mu}(x^{i-1})\right)$ 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 2.5.1.

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 code length 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 start-up problem. We note that this start-up problem is an inherent feature of the PML approach [26, 71], and our way of handling it is in line with the suggestions in [26].

2.3.6 Objective Bayesian Approach

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

$$P_{\mathrm{BAYES}}\left(x^n\right) = \int_{\Theta} P\left(x^n|\theta\right) w(\theta)\,\mathrm{d}\theta, \tag{2.8}$$

with universal code length $L_{\mathrm{BAYES}}\left(x^n\right) = -\ln P_{\mathrm{BAYES}}\left(x^n\right)$. Like NML, this can be approximated with an asymptotic formula. Under conditions similar to those for the NML approximation (2.1), we have [3]:

$$L_{\mathrm{ABAYES}}\left(x^n\right) := L\left(x^n|\hat{\theta}\right) + \frac{k}{2}\ln\frac{n}{2\pi} + \ln\frac{\sqrt{\det I(\theta)}}{w(\theta)}, \tag{2.9}$$

where the asymptotic behaviour is the same as for the approximation of the NML code length, roughly $L_{\text{ABAYES}}(x^n) - L_{\text{BAYES}}(x^n) \to 0$ as $n \to \infty$ (see below Eq. (2.1) 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 [3], which implies that the obtained results are independent of the parameterisation of the model [47]. It is defined as follows:

$$w(\theta) = \frac{\sqrt{\det I(\theta)}}{\int_\Theta \sqrt{\det I(\theta)}\,\mathrm{d}\theta}. \tag{2.10}$$

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 [14] for more information on objective Bayesian theory. The resulting universal codes with lengths $L_{\text{BAYES}}\left(x_2, \ldots, x_n \mid x_1\right)$ are, in fact, *conditional* on the first outcome. Recent work by [58] 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 does suggest 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 [4]. 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 [13], arbitrariness can be reduced by conditioning on every outcome in turn and then using the mean or median code length 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 $P_{\text{BAYES}}\left(x_2^n \mid x_1\right)$ for the Poisson and Geometric models. As before, the difference between the corresponding code lengths defines a model selection criterion. We also compute $P_{\text{ABAYES}}\left(x_2^n \mid x_1\right)$

for both models, the approximated version of the same quantity, based on approximation formula (2.9). 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:

$$w_{\mathrm{P}}(\mu) \propto \sqrt{I_{\mathrm{P}}(\mu)} = \mu^{-\frac{1}{2}}; \tag{2.11}$$

$$w_{\mathrm{P}}(\mu \mid x_1) = \frac{P_{\mathrm{P}}(x_1|\mu)\, w_{\mathrm{P}}(\mu)}{\int_0^\infty P_{\mathrm{P}}(x_1|\theta)\, w_{\mathrm{P}}(\theta)\, d\theta} = \frac{e^{-\mu}\mu^{x_1-\frac{1}{2}}}{\Gamma(x_1+\frac{1}{2})}. \tag{2.12}$$

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

$$P_{\mathrm{P,BAYES}}(x_2^n \mid x_1) = \int_0^\infty P_{\mathrm{P}}(x_2^n|\mu)\, w_{\mathrm{P}}(\mu \mid x_1)\, d\mu = \frac{\Gamma(s_n+\frac{1}{2})}{\Gamma(x_1+\frac{1}{2})} \Big/ \left(n^{s+\frac{1}{2}} \prod_{i=2}^{n} x_i! \right). \tag{2.13}$$

For the approximation (2.9) we obtain:

$$L_{\mathrm{P,ABAYES}}(x_2^n \mid x_1) = L_{\mathrm{P}}(x_2^n|\hat{\mu}(x_2^n)) + \tfrac{1}{2}\ln\frac{n}{2\pi} + \hat{\mu}(x_2^n) - x_1 \ln \hat{\mu}(x_2^n) + \ln\Gamma(x_1+\tfrac{1}{2}). \tag{2.14}$$

Bayesian code for the Geometric model We perform the same computations for the Geometric model. This time we get:

$$w_{\mathrm{G}}(\mu) \propto \mu^{-\frac{1}{2}}(\mu+1)^{-\frac{1}{2}}; \tag{2.15}$$

$$w_{\mathrm{G}}(\mu \mid x_1) = (x_1+\tfrac{1}{2})\mu^{x_1-\frac{1}{2}}(\mu+1)^{-x_1-\frac{3}{2}}; \tag{2.16}$$

$$P_{\mathrm{G,BAYES}}(x^n) = (x_1+\tfrac{1}{2})\frac{\Gamma(s+\frac{1}{2})\Gamma(n)}{\Gamma(n+s+\frac{1}{2})}. \tag{2.17}$$

For the approximation we obtain:

$$L_{\mathrm{G,ABAYES}}(x_2^n \mid x_1) = L_{\mathrm{G}}(x_2^n|\hat{\mu}(x_2^n)) + \tfrac{1}{2}\ln\frac{n}{2\pi} + x_1 \ln\left(1+\frac{1}{\hat{\mu}(x_2^n)}\right) + \tfrac{1}{2}\ln(\hat{\mu}(x_2^n)) - \ln(x_1+\tfrac{1}{2}). \tag{2.18}$$

2.4 Experiments

The previous section describes four methods 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 code length 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 code lengths 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 2.3, 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 code length 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 $L_{\mathrm{P}}\left(x^n|\mu\right) < L_{\mathrm{G}}\left(x^n|\mu\right)$. 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 [25], 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 error probability is exponentially decreasing with the same exponent [25, 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 2.2.

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.

2.4.1 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 2.2, 2.4, 2.5 and 2.6 plot the sample size against the error frequency, using different means μ and different priors p on the generating distribution. We use a log-scale, 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 2.4, 2.5 and 2.6 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 happens to work in the right direction), is thus treated as a baseline of sorts.

2.4.2 Bias

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

$$\Delta\left(x^n\right) := L_{\mathrm{G}}\left(x^n|\mu\right) - L_{\mathrm{P}}\left(x^n|\mu\right), \tag{2.19}$$

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

$$\Delta_{\mathrm{C}}(x^n) := L_{\mathrm{G,C}}\left(x^n\right) - L_{\mathrm{P,C}}\left(x^n\right) \tag{2.20}$$

(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,

$$E\left[\Delta_{\mathrm{C}}(X^n) - \Delta\left(X^n\right)\right]. \tag{2.21}$$

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 2.3.

2.4.3 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 code lengths and probability. If the universal code lengths 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 are 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 2.7. 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 [11]. At least in restricted contexts [23, 22], 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 [23].

2.5 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 PML and BIC/ML criteria 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 rather 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.

2.5.1 Poor Performance of the PML Criterion

One feature of Figure 2.2 that immediately attracts attention is the unusual slope of the error rate line of the PML criterion, which clearly favours the geometric distribution. This is even clearer in Figure 2.3, where the PML 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 PML criterion most severely underestimates the probability that the data were sampled from a Poisson distribution: of those sequences that were classified as geometric with 80% confidence, in fact about 60% turned out to be sampled from a Poisson distribution.

While this behaviour may seem appropriate if the data really are 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 p on the generating distribution, with nonzero probability for both distributions Poisson[μ] and Geom[μ]. The marginal

Figure 2.2 The $\log_{10}$ of the error frequency.

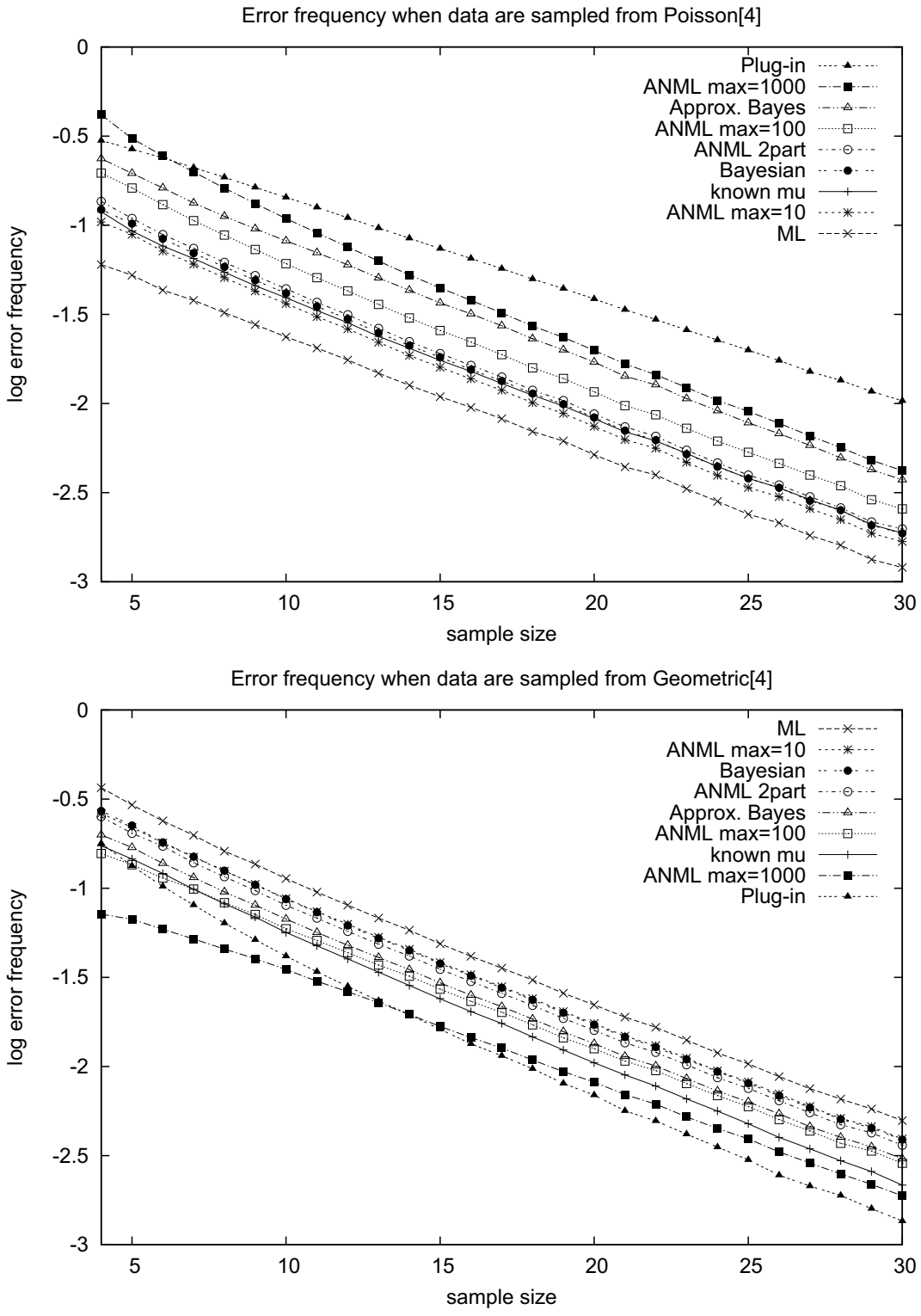


Figure 2.3 The classification bias in favour of the Poisson model in bits.

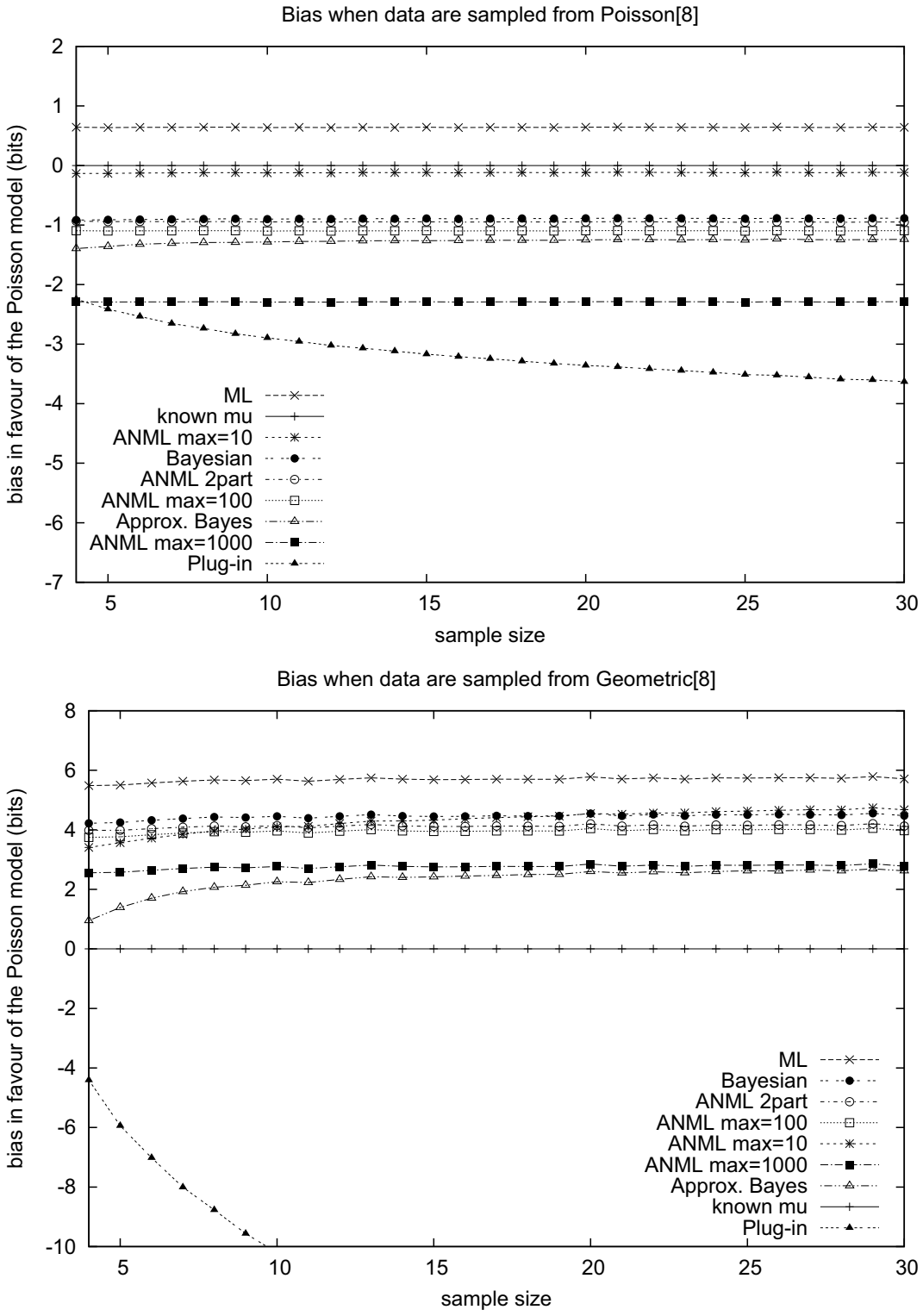


Figure 2.4 The difference in the $\log_{10}$ of the frequency of error between each criterion and the known μ criterion. The mean is 4. In these graphs, data are sampled from one of the two models with unequal probabilities.

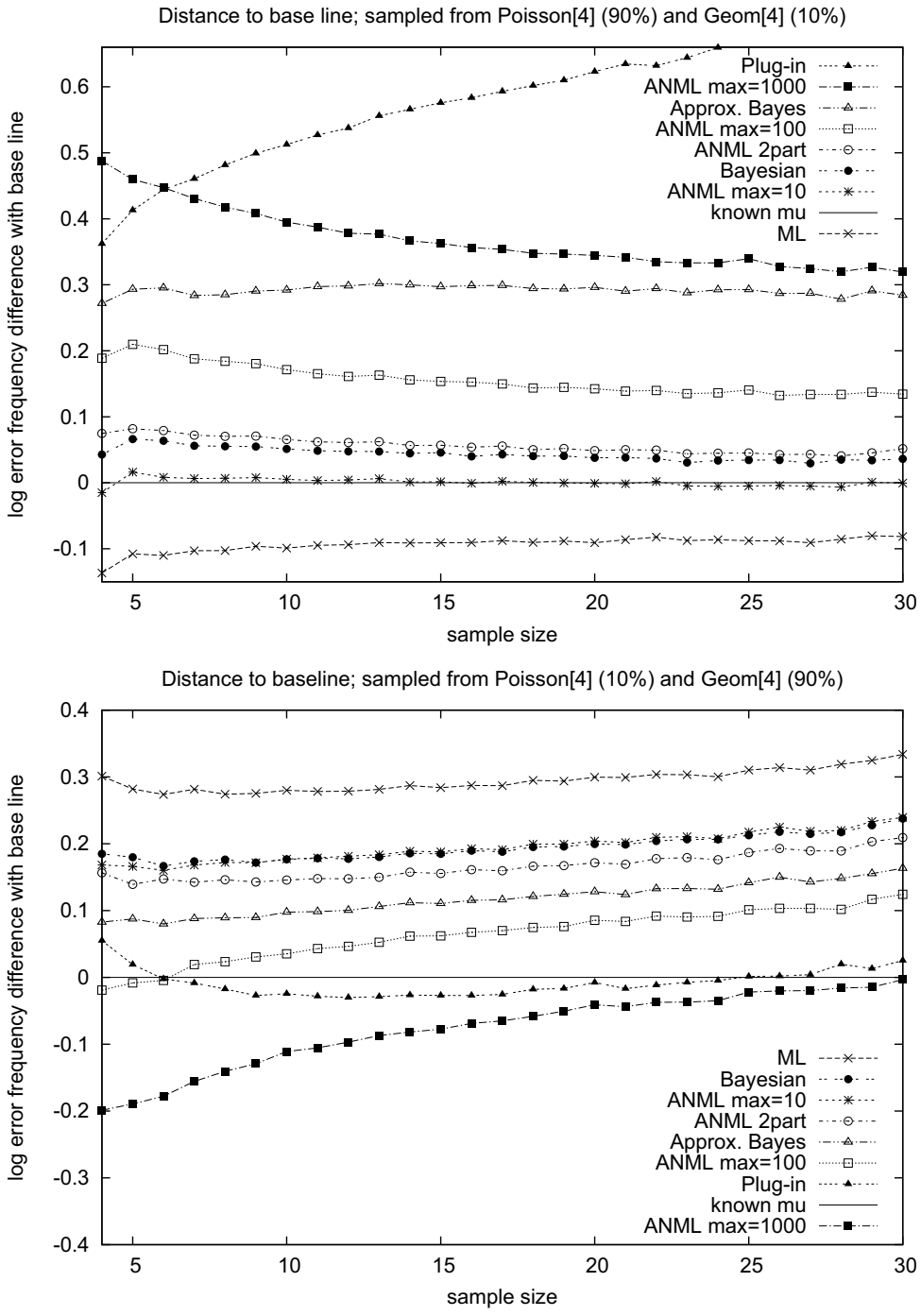


Figure 2.5 The difference in the $\log_{10}$ of the frequency of error between each criterion and the known μ criterion. The mean is 4.

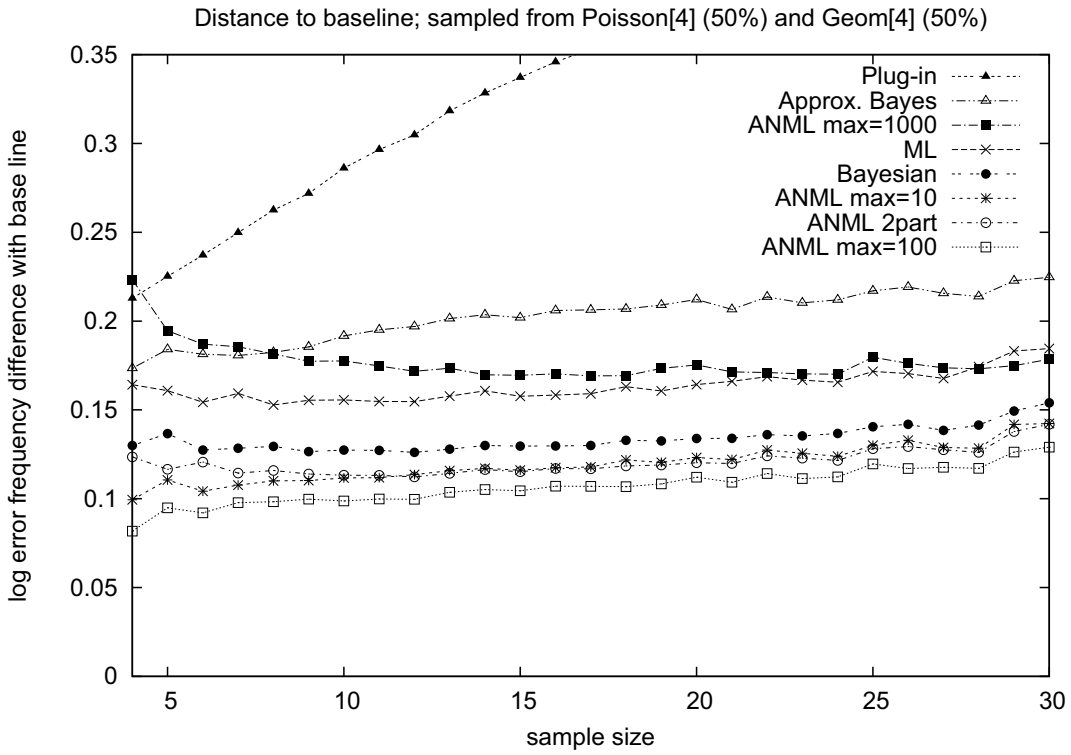
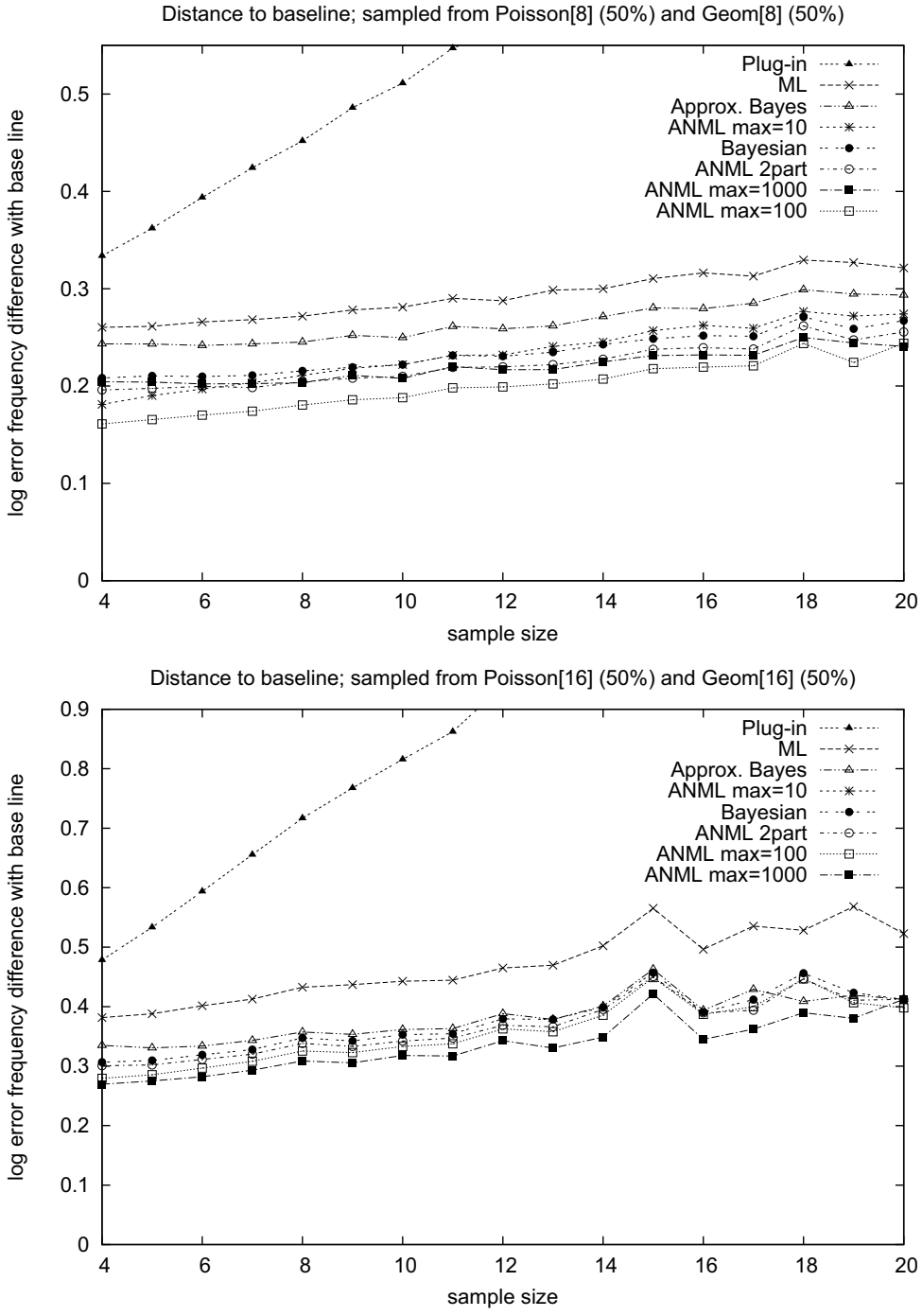


Figure 2.6 The difference in the $\log_{10}$ of the frequency of error between each criterion and the known μ criterion. The mean is 8 in the top graph and 16 in the bottom graph.



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 PML criterion is suboptimal. (As an aside, 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 2.4 shows that, even if there is a 90% chance that the data are geometric, then the PML criterion still has a worse (marginal) probability of error than "known μ" as soon as the sample size reaches 25. Figure 2.5 shows what happens if the prior on the generating distribution is uniform – using the PML 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 PML criterion initially came as a complete surprise to us. Theoretical literature certainly had not suggested it. Rather the contrary: in [71] we find that "it is only because of a certain inherent singularity in the process [of PML coding], as well as the somewhat restrictive requirement that the data must be ordered, that we do not consider the resulting predictive code length to provide another competing definition for the stochastic complexity, but rather regard it as an approximation". There are also a number of results to the effect that the regret for the PML code grows as $\frac{k}{2}\ln n$, the same as the regret for the NML code, for a variety of models. Examples are [70, 35, 97]. Finally, publications such as [61, 52] show excellent behaviour of the PML 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 PML 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 an estimate, $\hat{\mu}(x^{n-1})$ rather than the optimal μ to encode each outcome x_n is higher for the Poisson model. The accumulated difference accounts for the difference in regret.

This intuition is made precise in the next chapter, where we prove that for single parameter exponential families, the regret for the PML code grows with $\frac{1}{2}\ln(n)\mathrm{var}_{P^*}(X)/\mathrm{var}_{P_{\theta^*}}(X)$, where P^* is the generating distribution, while P_{θ^*} is the element of the model that minimises $D(P^*\|P_{\theta^*})$. The PML 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 $P^* = P_{\theta^*}$, so that the variances cancel.

Nevertheless, for some model selection problems, the PML criterion may be the best choice. For example, the Bayesian integral (2.8) cannot be evaluated analytically for many model families. It is then often approximated by, for example, Markov Chain Monte Carlo methods, and it is not at all clear whether the re-

sulting procedure will show better or worse performance than the PML criterion. Theoretical arguments [4] show that there are quite strong limits to how badly the PML criterion can behave in model selection. 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 PML 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 PML 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 PML 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.

All in all, our results indicate that the PML criterion should be used with caution, and may exhibit worse performance than other selection criteria under misspecification.

2.5.2 ML/BIC

Beside known μ and PML, 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 2.3 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 2.7 contains further evidence of bias in favour of the Poisson model: together with the PML 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 2.2. 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.

2.5.3 Basic Restricted ANML

We have seen that the ML/BIC criterion shows the largest bias for the Poisson model. Figure 2.3 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 2.3.2 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.

2.5.4 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.

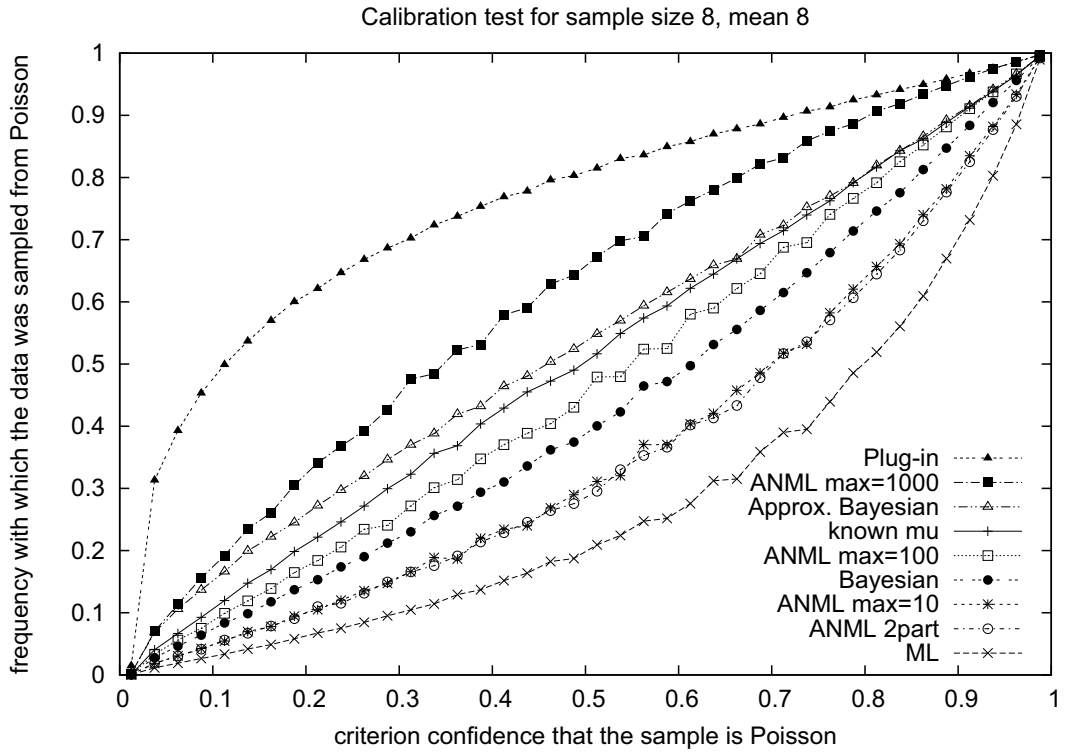
Figures 2.4–2.6 show that the error probability for these criteria tends to decrease at a slightly lower rate than for known μ (except when the prior on the generating distribution is heavily in favour of Poisson). While we do not understand this phenomenon well enough so as to prove it mathematically, 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.

2.6 Summary and Conclusion

We have experimented with a number of model selection criteria which are based on the MDL philosophy and involve computing the code length of the data with the help of the model. There are several ways to define such codes, 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

Figure 2.7 Calibration: probability that the model assigned to the data being Poisson against the frequency with which it actually was Poisson.



(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 PML code which always codes the new outcome using the distribution indexed by the maximum likelihood estimator for the preceding outcomes.

Only the NML code incurs the same regret for all possible data sequences x^n; the regret incurred by the alternatives we studied necessarily depends on the data. Arguably this dependence is quite weak for the objective Bayesian approach (see [39] for more details); how the regret depends on the data is at present rather unclear for the other approaches. As such they cannot really be viewed as "objective" alternatives to the NML code. However, new developments in the field make this distinction between "objective" and "subjective" codes seem a bit misleading. It is probably better to think in terms of "luckiness" (also see Section 1.1.1): while the NML code allows equally good compression (and thus learning speed) for all data sequences, other methods are able to learn faster from some data sequences than from others. This does not mean that conclusions drawn from inference with such luckiness codes are necessarily subjective.

We have performed error probability tests, bias tests and calibration tests to study the properties of the model selection criteria listed above and made the following observations.

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 chooses a neutral threshold, so the criterion is 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.

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 code length in an expected sense (Section 2.3.6).

The most surprising result is the behaviour of the PML criterion. It has a bias in favour of the Geometric model that depends strongly on the sample size. As a consequence, compared to the other model selection criteria its error rate decreases more slowly in the sample size if the data are sampled from each of the models with positive probability. This observation has led to a theoretical analysis of the code length of the PML code in the next chapter. It turns out that the regret of the PML 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.