MMLD Inference of the Poisson and Geometric Models

Daniel F. Schmidt and Enes Makalic

Clayton School of Information Technology, Monash University, Clayton, Victoria 3800, Australia

Abstract. This paper examines MMLD-based approximations to inference of two univariate probability densities: the geometric distribution and the Poisson distribution. The focus is on both parameter estimation and hypothesis testing properties of the approximation. The new parameter estimators are compared to the standard MML87 estimators in terms of bias, squared error loss and KL divergence. Empirical experiments demonstrate that the MMLD parameter estimates are more biased, and feature higher squared error loss than corresponding MML87 estimators. In contrast, the two criteria are virtually indistinguishable in the hypothesis testing experiment.

1 Introduction

Under the Minimum Message Length (MML) principle, inference is performed by seeking the model that admits the briefest encoding (i.e. most compression) of a two-part message sent from an imaginary sender to an imaginary receiver [1]. This paper considers the recent MMLD approximation [2] which in comparison to the well known MML87 [3] approximation has received little attention. In particular, we examine the problem of inference of two univariate exponential models, the Poisson and geometric distributions. The focus is on both parameter estimation and hypothesis testing properties of the approximation.

This paper is organised as follows. The MML87 and MMLD approximations are introduced in Section 2. Section 3 examines MMLD parameter estimation strategies. The Poisson and geometric distributions are examined in Sections 4 and 5 respectively. Hypothesis testing and concluding remarks are presented in Section 6 and Section 7.

2 Minimum Message Length Inference

2.1 The MML87 Approximation

The most popular MML approximation used in practice is the MML87 approximation [3]. Let $\mathbf{y}^n = (y_1, \ldots, y_n)$ denote n data samples, where $y_i \in \mathcal{Y} \subset \mathbb{R}$, and consider a model class $\mathcal{M}_\gamma$ of distributions $p(\cdot|\boldsymbol{\theta})$, indexed by a parameter vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$. The message length of data $\mathbf{y}^n$ given model $\boldsymbol{\theta}$ under the MML87 approximation is:

$$I_{87}\left(\mathbf{y}^n, \boldsymbol{\theta}, \boldsymbol{\gamma}\right) = -\log h(\boldsymbol{\theta}, \gamma) + \frac{1}{2}\log|\mathbf{J}(\boldsymbol{\theta}, \gamma)| - \log p(\mathbf{y}^n|\boldsymbol{\theta}, \gamma) + \frac{k}{2}\left(\log \kappa_k + 1\right) \quad (1)$$

where $h(\boldsymbol{\theta}, \gamma)$ is a prior distribution on the support Θ, $\mathbf{J}(\boldsymbol{\theta}, \gamma)$ is the Fisher information matrix, and κ_k is the normalised second moment of an optimal quantising lattice in k-dimensions. Define $\hat{\theta}_{87}(\mathbf{y}^n, \gamma)$ as the point estimate $\boldsymbol{\theta}$ that minimises (1) for model class $\mathcal{M}_\gamma$; this is henceforth referred to as the MML87 estimator relative to model class $\mathcal{M}_\gamma$. Inference is then performed by selecting the model that minimises the message length expression (1). In the sequel, the dependence on γ is dropped whenever it is clear from the context.

2.2 The MMLD Approximation

A recent alternative to the MML87 approximation that has received little attention in the literature is the MMLD approximation. Given data $\mathbf{y}^n$ and a set $\Omega \subset \Theta \subset \mathbb{R}^k$ the MMLD [4] message length is given by

$$I_{1D}\left(\mathbf{y}^n, \Omega\right) = -\log \int_\Omega h(\boldsymbol{\theta}) d\boldsymbol{\theta} - \frac{1}{\int_\Omega h(\boldsymbol{\theta}) d\boldsymbol{\theta}} \int_\Omega h(\boldsymbol{\theta}) \log p(\mathbf{y}^n|\boldsymbol{\theta}) d\boldsymbol{\theta} \qquad (2)$$

The set $\Omega(\mathbf{y}^n)$ that minimises the message length is deemed the *uncertainty region.* The uncertainty region is a sublevel set of the negative log-likelihood, $L(\cdot)$, and contains the distributions that are deemed indistinguishable on the basis of the available data and prior knowledge. This is similar in spirit to the concept of indistinguishable distributions [5] in the MDL literature [6, 7]. It is important to note that the MMLD message must be sent by the mechanism of random coding [1]. In general, the problem of finding $\Omega(\mathbf{y}^n)$ is computationally intractable, partly because of the difficulty in solving the integrals, and partly because of difficulties in determining the shape of the region. An important theorem is the following due to [1]:

Theorem 1. *Let $a(\boldsymbol{\theta})$ be a probability density over Θ, and $b(\boldsymbol{\theta})$ a continuous function of $\boldsymbol{\theta}$; the boundary of the set Ω that minimises*

$$-\log \int_\Omega a(\boldsymbol{\theta}) d\boldsymbol{\theta} + \frac{1}{\int_\Omega a(\boldsymbol{\theta}) d\boldsymbol{\theta}} \int_\Omega a(\boldsymbol{\theta}) b(\boldsymbol{\theta}) d\boldsymbol{\theta} \qquad (3)$$

lies on a level set of $b(\boldsymbol{\theta})$.

We assume that the likelihood function $p(\cdot)$ is everywhere continuous and non-redundant, i.e. that no two parameters θ index the same distribution. It is trivial to show that $\Omega(\mathbf{y}^n)$ satisfies the Boundary Rule by setting $a(\theta) = h(\theta)$ and $b(\theta) = -\log p(\mathbf{y}^n|\theta)$ in Theorem 1. Given the two assumptions about $p(\cdot)$, the result follows. The negative log-likelihood is continuous, strictly increasing around the single minimum $\hat{\theta}_{\mathrm{ML}}(\mathbf{y}^n)$. This implies that each level set of $L(\cdot)$

forms a closed curve. Since the boundary of $\Omega(\mathbf{y}^n)$ lies on a level set, the region must be contiguous. Given that $\Omega(\mathbf{y}^n)$ is a sublevel set of $L(\cdot)$, and $L(\hat{\theta}_{\text{ML}}(\mathbf{y}^n))$ is the infimum of $L(\cdot)$, it follows that $\hat{\theta}_{\text{ML}}(\mathbf{y}^n) \in \Omega(\mathbf{y}^n)$. Finally, the uncertainty region $\Omega(\mathbf{y}^n)$ is unique if the parameter space Θ is pruned such that $(\forall\theta \in \Theta, h(\theta) > 0)$.

3 MMLD Parameter Estimation

The uncertainty region $\Omega(\mathbf{y}^n)$ gives a set of possible models, but no indication how to choose one as representative. We define an MMLD parameter estimator as one that is restricted to select a $\boldsymbol{\theta} \in \Omega(\mathbf{y}^n)$. There exist various proposals for which a model can be chosen as the MMLD point estimate, but none has been adopted as the definitive standard. The estimators suggested in [2] are generalised by

$$\hat{\theta} = \arg \min_{\theta^* \in \Omega(\mathbf{y}^n)} \left\{ \int_{\Omega(\mathbf{y}^n)} \xi(\theta) \Delta(\theta || \theta^*) d\theta \right\} \tag{4}$$

where $\xi(\theta) \propto p(\mathbf{y}^n|\theta)h(\theta)$ (better for prediction) or $\xi(\theta) \propto h(\theta)$ (better for induction). Both estimators are, for most problems, difficult to compute (and must often be found numerically), and are a form of Bayesian decision estimator driven by a specific loss function (the Kullback-Leibler (KL) divergence [8]). This paper considers two alternate estimators that are much simpler to compute than (4), and are more in the spirit of the pure 'random coding' approach from which the MMLD approximation is derived. The first is the so called 'random coding' estimator, MMLD$_{\text{R}}$, which prescribes that the estimate simply be randomly picked from the prior over $\Omega(\mathbf{y}^n)$, $h_\Omega(\cdot)$. This is suggested in [1] (page 212) and is reasonable given that the uncertainty region represents the set of models that could be considered possible given the observed data. The 'random coding' estimator is also invariant under one-to-one model reparameterisations. It may be desirable to have an estimator that is non-randomised but based on the properties of $h_\Omega(\cdot)$. The obvious choice is to use the expected value of $\theta \in \Omega(\mathbf{y}^n)$, but this statistic is not invariant under one-to-one reparameterisations. Instead, the median estimator, MMLD$_{\text{MED}}$ is selected:

$$\hat{\theta}_{\text{MMLD}_{\text{MED}}}(\mathbf{y}^n) = \left\{ \theta \in \Omega(\mathbf{y}^n) : \int_a^\theta h_\Omega(x) dx = \frac{1}{2} \right\} \tag{5}$$

This estimator is easy to compute, at least for one-parameter models, and is invariant under one-to-one reparameterisations.

4 Poisson Distribution

The first distribution under consideration is the Poisson distribution. The probability of n i.i.d. data points $\mathbf{y}^n = (y_1, \ldots, y_n)$ generated by the Poisson distribution with rate parameter $\lambda > 0$ is:

$$p(\mathbf{y}^n|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \tag{6}$$

where $y_i \in \mathbb{Z}^+ \subset \mathcal{Y}$. Let $\bar{y} = \sum_{i=1}^{n} y_i$ be the sufficient statistic. The negative log-likelihood of data $\mathbf{y}^n$ given λ is:

$$L(\lambda) = -\log p(\mathbf{y}^n|\lambda) = n\lambda - \bar{y}\log\lambda + \sum_{i=1}^{n} \log y_i! \tag{7}$$

The maximum likelihood (ML) estimator is:

$$\hat{\lambda}_{\mathrm{ML}}(\mathbf{y}^n) = \frac{\bar{y}}{n} \tag{8}$$

The MML principle requires a suitable prior density defined over all the model parameters. This paper uses the conjugate exponential distribution:

$$h(\lambda|\alpha) = \alpha e^{-\alpha\lambda} \tag{9}$$

where $\alpha > 0$; as $\alpha \to \infty$ this prior places more probability mass on smaller λ values, and as $\alpha \to 0$ it becomes increasingly diffuse. The message length of $\mathbf{y}^n$ given λ under the MML87 approximation is:

$$I_{87}(\mathbf{y}^n, \lambda) = (n+\alpha)\lambda - \left(\bar{y} + \frac{1}{2}\right)\log\lambda + \frac{1}{2}\log\left(\frac{n}{12\alpha^2}\right) + \sum_{i=1}^{n}\log y_i! + \frac{1}{2} \tag{10}$$

The estimate $\hat{\lambda}_{87}(\mathbf{y}^n)$ that minimises (10) is:

$$\hat{\lambda}_{87}(\mathbf{y}^n) = \frac{\bar{y} + \frac{1}{2}}{n+\alpha} \tag{11}$$

In comparison to the ML estimate, the inclusion of the prior term extends the data by α artificial counts of zero. The effect of the Fisher information term is to increase the total sum of the counts by $\frac{1}{2}$. While this introduces a bias into the estimate, even in the case of $\alpha = 0$ (i.e. a uniform prior), it rules out an inadmissible estimate of $\hat{\lambda}(\mathbf{y}^n) = 0$.

4.1 Results

The results of parameter estimation experiments are now presented. The three estimators (MMLD$_{\mathrm{R}}$, MMLD$_{\mathrm{MED}}$ and MML87) are compared in terms of their bias, squared error loss and KL divergence. The exact expressions for the MML87 bias and squared error loss are derived, and for the MMLD parameter estimators the distribution of the sufficient statistic is exploited to efficiently compute the required expectations. The prior hyperparameter is set to $\alpha = 0.001$ for all tests. This leads to a prior that is almost uniform over the range of λ_* under consideration, and means the estimators are not heavily influenced by the effect of the prior. The results are presented in Table 1. It is immediately clear that the random coding estimator performs uniformly worse in comparison to the $\mathrm{MMLD}_{\mathrm{MED}}$ and MML87 estimators on all three metrics. The bias of the $\mathrm{MMLD}_{\mathrm{R}}$ and $\mathrm{MMLD}_{\mathrm{MED}}$ estimators is virtually identical. This is not surprising as the mean and median of the truncated exponential distribution almost coincide. Interestingly, although the bias of the $\mathrm{MMLD}_{\mathrm{MED}}$ estimator is uniformly greater than that of the MML87 estimator, its squared error loss is not significantly worse. In terms of KL divergence, the $\mathrm{MMLD}_{\mathrm{MED}}$ estimator is uniformly better in comparison to the MML87 estimator.

Table 1. Poisson parameter estimation results

		$\mathrm{MMLD}_{\mathrm{R}}$			$\mathrm{MMLD}_{\mathrm{MED}}$			MML87		
n	λ_*	BIAS	SPE	KL	BIAS	SPE	KL	BIAS	SPE	KL
5	1	0.1992	0.4472	0.1934	0.1990	0.2396	0.0951	0.0998	0.2099	0.1021
	5	0.1993	2.0455	0.1995	0.1988	1.0377	0.0991	0.0990	1.0094	0.1003
	20	0.1957	8.0415	0.1998	0.1937	4.0333	0.0997	0.0960	4.0076	0.1001
10	1	0.1006	0.2132	0.0998	0.1006	0.1103	0.0491	0.0499	0.1025	0.0505
	5	0.0993	1.0115	0.0999	0.0990	0.5092	0.0498	0.0495	0.5023	0.0501
	20	0.0980	4.0107	0.1000	0.0970	2.0084	0.0499	0.0480	2.0019	0.0500
25	1	0.0404	0.0829	0.0405	0.0404	0.0421	0.0201	0.0200	0.0404	0.0201
	5	0.0399	0.4030	0.0401	0.0398	0.2016	0.0200	0.0198	0.2004	0.0200
	20	0.0394	1.6019	0.0400	0.0390	0.8015	0.0200	0.0192	0.8003	0.0200
50	1	0.0212	0.0413	0.0203	0.0212	0.0209	0.0102	0.0100	0.0201	0.0100
	5	0.0193	0.2008	0.0200	0.0192	0.1002	0.0100	0.0099	0.1001	0.0100
	20	0.0193	0.8004	0.0200	0.0191	0.4005	0.0100	0.0096	0.4001	0.0100

The second test undertaken was to compare the MMLD message length against the MML87 message length for the same data sets. Recall that the MMLD message length is constructed explicitly to minimise transmission by random coding. Since the MML87 message must also be transmitted by the device of random coding, it follows that $I_{87}(\mathbf{y}^n) \geq I_D(\mathbf{y}^n)$. Table 2 presents the fractional errors of the MML87 code length relative to the MMLD code length, for a range of α and n values. It can be seen that the mean and maximum approximation error of the MML87 approximation in general increases with smaller n and larger α. This is not surprising since for smaller n the assumption of locally quadratic negative log-likelihood curves may be violated. As α increases, the curvature of the prior density over the MML87 uncertainty region becomes increasingly larger, which leads to a substantial violation of the assumption of a locally flat prior. The maximum errors were obtained when $\bar{y}$ was close to zero, and as $\bar{y} \to \infty$ the approximation grew increasingly better. This is due to the fact that as $\bar{y}$ increases, so does $\hat{\lambda}_{87}(\mathbf{y}^n)$. For larger $\hat{\lambda}_{87}(\mathbf{y}^n)$, the Fisher Information is smaller and the exponential prior becomes increasingly flat.

Table 2. Fractional differences between MMLD and MML87 message lengths for the Poisson model

	$\alpha = 0.01$		$\alpha = 0.1$		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 2$	
n	mean	max	mean	max	mean	max	mean	max	mean	max
5	0.0010	0.0311	0.0016	0.0487	0.0036	0.0821	0.0069	0.1160	0.0144	0.1858
10	0.0009	0.0281	0.0012	0.0416	0.0020	0.0633	0.0031	0.0821	0.0058	0.1160
25	0.0007	0.0249	0.0010	0.0349	0.0014	0.0487	0.0017	0.0590	0.0024	0.0750
50	0.0007	0.0230	0.0009	0.0311	0.0011	0.0416	0.0013	0.0487	0.0016	0.0590

5 Geometric Distribution

The next distribution under consideration is the geometric distribution. Following the lead of [9], we parametrise the geometric distribution in terms of a mean parameter $\mu > 0$:

$$p(\mathbf{y}^n|\mu) = \prod_{i=1}^{n} \frac{\mu^{y_i}}{(\mu+1)^{y_i}} \tag{12}$$

where $y_i \in \mathbb{Z}^+ \subset \mathcal{Y}$. The geometric and Poisson distributions share the same sufficient statistic. The negative-log likelihood of data $\mathbf{y}^n$ given μ is:

$$L(\mu) = -\log p(\mathbf{y}^n|\mu) = n\log(\mu+1) - \bar{y}\log\left(\frac{\mu}{\mu+1}\right) \tag{13}$$

To allow for easy comparison to the Poisson distribution in the sequel, we place an exponential prior over the mean. The message length of $\mathbf{y}^n$ given μ under the MML87 approximation is:

$$I_{87}(\mathbf{y}^n, \mu) = \left(n - \bar{y} - \frac{1}{2}\right)\log\mu + \left(\bar{y} - \frac{1}{2}\right)\log(\mu+1) + \alpha\mu + \frac{1}{2}\log\left(\frac{n}{12\alpha^2}\right) + \frac{1}{2} \tag{14}$$

The estimate $\hat{\mu}_{87}(\mathbf{y}^n)$ that minimises (14) is the positive solution to the quadratic equation:

$$\alpha\mu^2 + (n+\alpha-1)\mu + \left(n - \bar{y} - \frac{1}{2}\right) = 0 \tag{15}$$

5.1 Results

The results of the geometric parameter estimation tests are presented in Table 3. The MML87 estimator performs consistently better than the MMLD estimators in terms of all three metrics considered. The MMLD$_R$ estimator performs consistently worse than the MMLD$_{\text{MED}}$ estimator, though both have similar levels of bias. The KL divergences of the MMLD$_{\text{MED}}$ estimator are close to the MML87 estimator, though its squared prediction error is naturally larger than the MML87 estimator given it has a larger bias. It is known that the MML87 estimator coincides with the ML estimator asymptotically, and it is conjectured

Table 3. Geometric parameter estimation results

		$\mathrm{MMLD}_{\mathrm{R}}$			$\mathrm{MMLD}_{\mathrm{MED}}$			MML87		
n	λ_*	BIAS	SPE	KL	BIAS	SPE	KL	BIAS	SPE	KL
5	1	-0.7840	2.4103	0.2057	-0.7836	1.5379	0.1286	-0.3740	0.7635	0.1145
	5	-2.8585	34.777	0.2258	-2.8521	21.884	0.1368	-1.3610	11.150	0.1197
	20	-10.471	476.218	0.2250	-10.382	294.32	0.1350	-4.9315	151.67	0.1186
10	1	-0.3402	0.7084	0.1049	-0.3400	0.4163	0.0583	-0.1664	0.2744	0.0550
	5	-1.2432	10.385	0.1061	-1.2410	6.0365	0.0590	-0.6066	4.0606	0.0545
	20	-4.5808	144.05	0.1058	-4.5505	82.916	0.0586	-2.2148	56.186	0.0543
25	1	-0.1259	0.2027	0.0409	-0.1259	0.1098	0.0214	-0.0624	0.0907	0.0207
	5	-0.4607	3.0089	0.0409	-0.4600	1.6201	0.0214	-0.2278	1.3527	0.0207
	20	-1.7013	41.951	0.0409	-1.6915	22.471	0.0214	-0.8334	18.832	0.0206
50	1	-0.0614	0.0902	0.0202	-0.0614	0.0471	0.0103	-0.0306	0.0426	0.0102
	5	-0.2248	1.3456	0.0202	-0.2244	0.7001	0.0104	-0.1116	0.6369	0.0102
	20	-0.8305	18.8014	0.0202	-0.8260	9.7610	0.0103	-0.4093	8.8986	0.0102

that any MMLD estimator will asymptotically converge to the Maximum Likelihood estimator as $n \to \infty$. The results support this as they demonstrate that as n increases, the differences between the estimators begins to diminish.

A comparison of the MML87 and MMLD code lengths for the geometric distribution is given in Table 4. The results are quite similar to those obtained in the test of the Poisson distribution. As before, a decrease in n or increase in α results in a degradation of the MML87 approximation, leading to larger approximation errors. The maximum approximation errors obtained for the geometric distribution are lower in comparison to those obtained for the Poisson distribution, for the same values of n and α. However, the mean approximation error for the geometric distribution is generally larger. Given both models are using the same prior, these differences must be attributed to the Fisher Information term. On average, the quadratic likelihood approximation is less valid for the geometric distribution than the Poisson, but in the worst case the quadratic approximation becomes worse for the Poisson.

Table 4. Fractional differences between MMLD and MML87 message lengths for the geometric model

	$\alpha = 0.01$		$\alpha = 0.1$		$\alpha = 0.5$		$\alpha = 1$		$\alpha = 2$	
n	mean	max	mean	max	mean	max	mean	max	mean	max
5	0.0023	0.0222	0.0024	0.0356	0.0168	0.0647	0.0301	0.0974	0.0452	0.1694
10	0.0011	0.0242	0.0011	0.0360	0.0040	0.0559	0.0089	0.0737	0.0177	0.1070
25	0.0007	0.0236	0.0009	0.0330	0.0013	0.0463	0.0021	0.0562	0.0038	0.0718
50	0.0006	0.0223	0.0008	0.0303	0.0011	0.0405	0.0013	0.0475	0.0018	0.0576

6 Hypothesis Testing

This section examines the application of the MMLD approximation to hypothesis testing. In particular, we consider the problem previously studied in the context of MDL [9]. The task is identify whether a set of discrete counts was generated by a Poisson distribution or a geometric distribution. In the MML framework this amounts to finding the message lengths under both hypotheses and selecting the one that yields the lower message length. The MML approximations are compared against the Normalised Maximum Likelihood (NML) universal model [6] which encodes data $\mathbf{y}^n$ using model class $\mathcal{M}_\gamma$ with code length

$$SC(\mathbf{y}^n) = -\log p(\mathbf{y}^n|\hat{\theta}_{\rm ML}(\mathbf{y}^n), \gamma) + \log \int_{\mathbf{x}^n \in \mathcal{Y}} p(\mathbf{x}^n|\hat{\theta}_{\rm ML}(\mathbf{x}^n), \gamma) d\mathbf{x}^n \tag{16}$$

Under suitable regularity conditions an asymptotic approximation to (16) is given by [10]. In comparison to the MML two-part codes, this code possesses the property of minimax optimality with respect to worst-case regret.

Table 6 presents results of hypothesis testing for several values of n and α. The MML approximations are compared against the regular NML code (NML1) on the limited parameter range $\theta \in (0, 1000)$, and the two-part NML code (NML2) proposed in [9]. The criteria are assessed on their 0/1 classification loss and the model class selection 'bias'. The latter is the frequency at which the criteria selected the geometric distribution. Given that the data was generated with equal probability from either a Poisson or a geometric source, an unbiased criterion is expected to have a 'bias' score of roughly 0.5. For each value of n and α, the experiment was repeated 10^5 times.

Table 5. Average message lengths for $n = 3$

	Poisson				Geometric			
α	MMLD	MML87	Difference	Fract. Diff	MMLD	MML87	Difference	Fract. Diff.
0.01	53.656	53.653	0.0027	0.0001	16.289	16.195	0.0924	0.0057
1	4.7578	4.6294	0.1284	0.0270	4.7260	4.5135	0.2125	0.0450

For almost all values of n and α, the two NML criteria performed marginally worse than the MML criteria. The most notable difference arises for $n = 3$ and $\alpha = 1$. In this case, the NML1 criterion is quite heavily biased towards the geometric distribution. These results corroborate the findings in [9]. Of further interest is the difference in bias between the MMLD and MML87 criteria for $n = 3$ and $\alpha = 1$. Examination of Tables 2 and 4 shows that it is exactly for these conditions that the MML87 message length approximation breaks down. The average message lengths of the geometric and Poisson distributions for both the MMLD and MML87 approximations are shown in Table 5. It is clear that for $\alpha = 1$, when the sufficient statistic $\bar{y}$ is more likely to be small, the MML87 approximation underestimates the geometric code length more than it underestimates the Poisson code length. Given the tiny amounts of data, this has significant effect on the selection bias. In contrast, when $\alpha = 0.01$ the underestimation errors are insignificant and do not affect the selection bias.

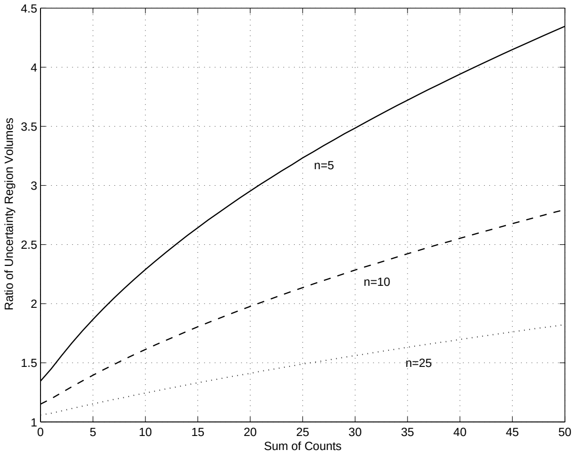

Fig. 1. Ratio of the volume of the uncertainty regions of the geometric and Poisson distributions

A final test was to examine the behaviour of the uncertainty regions of the two distributions across different n and sufficient statistics $\bar{y}$. Figure 1 shows the ratio of the volume of the geometric uncertainty region $\Omega_G(\bar{y})$ to the volume of the Poisson uncertainty region $\Omega_P(\bar{y})$ for the same sufficient statistics $\bar{y}$. The plots demonstrate that the volume of $\Omega_G(\bar{y})$ is always larger than the volume of $\Omega_P(\bar{y})$, and increases as $\bar{y}$ grows. As n increases, the ratio becomes smaller, which is anticipated given that asymptotically the two uncertainty regions coincide. This is in line with the findings in [9] which show that the parametric complexity of the Poisson is always greater than that of the geometric distribution.

Table 6. Hypothesis testing results

		MMLD		MML87		NML1		NML2	
n	α	Correct	Bias	Correct	Bias	Correct	Bias	Correct	Bias
3	1	0.5895	0.2354	0.5895	0.6236	0.5059	0.7982	0.5882	0.2223
	0.01	0.9383	0.4873	0.9380	0.4870	0.9313	0.5161	0.9330	0.4554
9	1	0.6730	0.5103	0.6732	0.5134	0.5891	0.8909	0.6692	0.4973
	0.01	0.9889	0.4989	0.9889	0.4988	0.9862	0.5099	0.9884	0.4936
21	1	0.7492	0.4797	0.7491	0.4801	0.6685	0.8116	0.7489	0.4968
	0.01	0.9951	0.4988	0.9950	0.4988	0.9934	0.5054	0.9951	0.4975

7 Conclusion

This paper has applied the MMLD approximation to the problem of parameter estimation and model selection of Poisson and geometric distributions. In particular, it examined parameter estimation schemes based on the notion of random coding and compared them to the corresponding MML87 estimators. We introduced a new MMLD parameter estimation method and examined the bias and squared error loss properties of the MMLD and MML87 estimators. Additionally, the ability of the MML-based criteria to perform hypothesis testing was compared against the latest MDL approximation.

The main findings of the paper are as follows. The random coding estimator $\mathrm{MMLD}_{\mathrm{R}}$ appears inferior to the other MML estimators that pick a single non-randomised estimate. The median estimator $\mathrm{MMLD}_{\mathrm{MED}}$ is almost comparable to the MML87 estimator in terms of squared error loss and KL divergence, but remains well defined even when the MML87 estimator produces nonsensical results. Hence, $\mathrm{MMLD}_{\mathrm{MED}}$ may be considered a robust alternative to MML87 for difficult model classes. Given the good performance of the MML87 estimators, it is conjectured that an MMLD-like approximation based on the expected likelihood is a promising avenue of future research. Although all criteria considered obtained similar classification results, the positive MML87 performance for small n and large α is possibly more through accident than design.

References

1. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. First edn. Information Science and Statistics. Springer (2005)
2. Fitzgibbon, L.J.: Message From Monte Carlo: A Framework for Minimum Message Length Inference using Markov Chain Monte Carlo Methods. PhD thesis, Faculty of Information Technology, Monash University (2004)
3. Wallace, C.S., Freeman, P.R.: Estimation and inference by compact coding. J. Royal Statistical Society B **49** (1987) 240–252
4. Fitzgibbon, L.J., Dowe, D.L., Allison, L.: Message from Monte Carlo. Tech Report 2002/107, School of Computer Science and Software Engineering, Monash University, Australia 3800 (Decemember 2002)
5. Balasubramanian, V.: MDL, Bayesian inference, and the geometry of the space of probability distributions. In P. D. Grünwald, I.J.M., Pitt, M.A., eds.: Advances in Minimum Description Length: Theory and Applications. MIT Press (2005) 81–99
6. Rissanen, J.: Information and Complexity in Statistical Modeling. First edn. Information Science and Statistics. Springer (2007)
7. Grünwald, P.D.: The Minimum Description Length Principle. Adaptive Communication and Machine Learning. The MIT Press (2007)
8. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics **22**(1) (March 1951) 79–86
9. de Rooij, S., Grünwald, P.: An empirical study of minimum description length model selection with infinite parametric complexity. Journal of Mathematical Psychology **50**(2) (April 2006) 180–192
10. Rissanen, J.: Fisher information and stochastic complexity. IEEE Transactions on Information Theory **42**(1) (1996) 40–47