A NEW MESSAGE LENGTH APPROXIMATION FOR PARAMETER ESTIMATION AND MODEL SELECTION

Daniel F. Schmidt[1]

[1]Centre for MEGA Epidemiology, The University of Melbourne,
Melbourne, AUSTRALIA, dschmidt@unimelb.edu.au

ABSTRACT

This paper examines Bayesian two-part coding schemes as tools for parameter estimation and model selection. The Wallace–Freeman message length approximation to strict minimum message length can be used to obtain two-part message lengths. However, this approximation relies on some strong assumptions regarding the likelihood function and prior distribution which do not hold for a large range of models. We present a new two-part message length formula that is more widely applicable than the popular Wallace–Freeman message length approximation, while remaining significantly easier to compute than the exact strict minimum message length procedure.

1. MML TWO-PART CODES

Consider the problem of choosing a plausible explanation for some observed data $\mathbf{y}^n = (y_1, \ldots, y_n)' \in \mathcal{Y}^n \subseteq \mathbb{R}^n$. The possible explanations are the distributions, or (fully specified) models, contained in a countable set of parametric model structures $\gamma \in \Gamma$. Let $p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})$ denote the model[1], in model structure γ, indexed by $\boldsymbol{\theta} \in \Theta_\gamma \subseteq \mathbb{R}^k$. The minimum encoding approach [1, 2] to inference suggests that the model that most compresses the data is the most plausible explanation. One way to compress the data is by two-part coding, in which the model and the data are compressed together as a two-part message. This idea is central to the minimum message length principle (MML). The MML principle is explicitly Bayesian in nature, so we further assume that a suitable prior distribution, $\pi_\gamma(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta_\gamma$ exists for all $\gamma \in \Gamma$.

The first part of the message, or *assertion*, states which model, $p_\gamma(\cdot|\boldsymbol{\theta})$, from the structure γ, is to be used to compress the data. The second part, or *detail*, states the data $\mathbf{y}^n$ using the nominated model from γ. Let the length of these two terms be denoted by $I(\boldsymbol{\theta};\gamma)$ and $I(\mathbf{y}^n|\boldsymbol{\theta};\gamma)$, respectively. Further, let $I(\gamma)$ denote the length of a preamble code stating which structure from Γ is being used. Estimation of both a model structure, as well as the model parameters, may be simultaneously performed by solving

$$\left\{\hat{\gamma}, \hat{\boldsymbol{\theta}}\right\} = \mathop{\arg\min}_{\gamma\in\Gamma, \boldsymbol{\theta}\in\Theta_\gamma} \left\{ I(\gamma) + I(\boldsymbol{\theta};\gamma) + I(\mathbf{y}^n|\boldsymbol{\theta};\gamma) \right\}.$$

In the strict MML (SMML) procedure, the assertion and detail codes are constructed so that for a given structure γ, the expected joint codelength is minimised, the expectation being taken with respect to the marginal distribution of the data. The optimisation problem implicit in this minimisation is in general NP-hard [3], and thus the procedure is impractical for all but the simplest of problems.

1.1. The Wallace–Freeman Codelength

Under suitable regularity conditions, Wallace and Freeman proposed an approximate codelength formula which we shall refer to as MML87 [4]. For a structure γ with k free parameters, the MML87 assertion and detail lengths for a model $\boldsymbol{\theta} \in \Theta_\gamma$ are

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

$$I_{87}(\mathbf{y}^n|\boldsymbol{\theta};\gamma) = -\log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}) + \frac{k}{2}, \quad (2)$$

where

$$\mathbf{J}_\gamma(\boldsymbol{\theta}^*) = -\mathrm{E}_{\boldsymbol{\theta}^*}\left[\left.\frac{\partial^2 \log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right]$$

is the Fisher information matrix and κ_k is the normalised second moment of an optimal quantising lattice in k dimensions. Wallace has shown that if the local curvature of the prior distribution $\pi_\gamma(\cdot)$ is "small" in comparison to the curvature of the negative log-likelihood, the MML87 codelength (1–2) is virtually indistinguishable (pp. 230–231, [1]) from the exact SMML codelengths.

The Wallace–Freeman approximation is computationally tractable. However, the accuracy of the approximation depends crucially on the behaviour of the likelihood function and the prior. If the Fisher information matrix is near singular, or the curvature of the prior is too great, the MML87 codelength can be a poor approximation to the SMML codelength. This paper introduces a new two-part codelength formula, named "MML08" after the year of its introduction [5], that is robust to these problems, while remaining significantly easier to compute than the exact SMML codes.

2. RANDOM CODING AND MMLD

We now discuss the MMLD approximation [6] that was specifically proposed to provide a more robust alternative

[1] We acknowledge that this use of the term "model" differs from much of the traditional statistical literature. This is done to keep the terminology in this paper consistent with the MML literature.

to the MML87 approximation, and which forms the basis for the new MML08 codelength presented in Section 3. We provide a derivation of the MMLD codelength that differs from the one in [1] (pp. 210–213), and discuss how data may be transmitted using a model structure γ without the need to perform a complete discretization of the parameter space Θ_γ by using Wallace's ingenious procedure of *random coding*.

2.1. Random Coding

To transmit data $\mathbf{y}^n$ via random coding it is required that both the receiver and transmitter have access to a pseudo random number generator capable of sampling from the prior $\pi_\gamma(\cdot)$, and that both generators are initialized with the same seed. The transmitter repeatedly samples models from the prior distribution until they generate one that lies inside a set $S \subseteq \Theta_\gamma$. The transmitter then sends the number of draws required to arrive at the model, say d, to the receiver using a universal code for the integers, with codelength $l^*(d)$. This is the message assertion. The receiver then makes d draws from their random number generator to arrive at the same parameter vector. The transmitter may then use this model, say $\boldsymbol{\theta}_d$, to send the data; this is the detail of the message. The total message length is then

$$I(\mathbf{y}^n, d, \boldsymbol{\theta}_d; \gamma) = l^*(d) - \log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}_d).$$

The length of the code required to transmit a string $\mathbf{y}^n$ using random coding is a random variable that depends crucially on the choice of S. One wishes for the messages to be short on average and so S is chosen to minimise the average expected random coding message length, i.e.,

$$\underset{S \subseteq \Theta_\gamma}{\arg\min} \left\{ \mathrm{E}\left[l^*(d) - \log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}_d) \right] \right\},$$

where the expectation is taken with respect to the random variables $(d, \boldsymbol{\theta}_d)$. The MMLD message length is found by approximating $\mathrm{E}\left[l^*(d)\right]$ and then solving for the minimising set S. This is detailed in the next section.

2.2. MMLD and Average Codelengths

Observe that the random variables d and $\boldsymbol{\theta}_d$ are independent; it thus suffices to find the expectations for both components of the random coding message length individually. Let

$$q_\gamma(S) = \mathbb{P}(\boldsymbol{\theta} \in S) = \int_S \pi_\gamma(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

be the probability that a model $\boldsymbol{\theta}$ sampled randomly from $\pi_\gamma(\cdot)$ lies in S. The number of draws, d, required for a model to fall in S is a random variable following a geometric distribution with parameter $q_\gamma(S)$. To transmit d to the receiver we use a universal code for integers, such as the log-star code [2], or Wallace tree code [1]. The log-star codelength for integer d is

$$l^*(d) = \log d + \log\log d + \ldots$$

where the iterated logarithms continue until they become negative. Given that $\mathrm{E}\left[d\right] = 1/q_\gamma(S)$ and $\mathrm{var}(d) = (1 - q_\gamma(S))/q_\gamma(S)^2$, we can use the approximation $\mathrm{E}\left[l^*(d)\right] = \log 1/q_\gamma(S) + O(\log\log 1/q_\gamma(S))$. Using only the dominant term we arrive at the expression for the average length of the assertion of a random coding message based on the set S, $I(S;\gamma) = \log 1/q_\gamma(S)$. It remains to determine the average length of the detail. The distribution of $\boldsymbol{\theta}_d$, i.e., the first randomly generated model to lie in S, is

$$p(\boldsymbol{\theta}_d) = \frac{\pi_\gamma(\boldsymbol{\theta}_d)}{\int_S \pi_\gamma(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \quad \boldsymbol{\theta}_d \in S,$$

so that the average detail length of a message based on the set S is given by

$$I(\mathbf{y}^n|S;\gamma) = -\frac{1}{q_\gamma(S)} \int_S \pi_\gamma(\boldsymbol{\theta}) \log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (3)$$

The average total random coding message based on S is $I(\mathbf{y}^n, S; \gamma) = I(S;\gamma) + I(\mathbf{y}^n|S;\gamma)$. It is informative to define the "round-off" error as

$$r_\gamma(\mathbf{y}^n, S) = I(\mathbf{y}^n|S;\gamma) + \log p_\gamma(\mathbf{y}^n|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}), \quad (4)$$

where $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta} \in \Theta_\gamma} \{p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})\}$ is the maximum likelihood estimate. The quantity $r_\gamma(\cdot, S)$ can be interpreted as the increase in the length of the detail over the "maximum-likelihood" code incurred by using a quantised estimate, represented by S, in place of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, to transmit the data. Let $\Omega_\gamma(\mathbf{y}^n)$ denote the set that solves

$$\min_{S \subseteq \Theta_\gamma} \left\{ I(S;\gamma) - \log p_\gamma(\mathbf{y}^n|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) + r_\gamma(\mathbf{y}^n, S) \right\}. \quad (5)$$

In the MML literature, this set is called the *uncertainty region*, as it includes all models that are considered to be plausible explanations of the data. The MMLD codelength is then given by $I_{\mathrm{D}}(\mathbf{y}^n;\gamma) \equiv I(\mathbf{y}^n, \Omega_\gamma(\mathbf{y}^n);\gamma)$. Examining the minimisation problem (5) shows that the MMLD codelength can be interpreted as balancing the accuracy to which maximum likelihood estimates are stated against the evidence in the data.

Unfortunately, direct replacement of MML87 by the MMLD approximation is not possible. The MMLD codelength may be used to select a model structure γ, but offers no guidance for selection of a suitable point estimate. This is because the MMLD messages are essentially (redundant) one-part codes. The random coding procedure on which they are based is in theory two-part; however, the MMLD procedure, by integrating out the random variables $(d, \boldsymbol{\theta}_d)$ to arrive at a sensible measure of message length, removes the ability to transmit the data using an arbitrary model from Θ_γ. The data alone determines the uncertainty region $\Omega_\gamma(\mathbf{y}^n)$, and in this sense the MMLD message length offers a codebook over $\mathcal{Y}^n$ only. The next section proposes an new message length formula that addresses this issue.

3. THE MML08 CODELENGTH

The main contribution of this paper is to present a generalization of the MMLD message length equation that allows one to derive point estimates explicitly by minimising the joint model and data codelength. In this way it acts as a replacement for MML87 when the Wallace–Freeman assumptions do not apply, and is significantly easier to compute than the exact SMML codelength.

3.1. Model Cost

Define the quantity

$$D_D(\mathbf{y}^n;\gamma) = \log 1/q_\gamma\left(\Omega_\gamma(\mathbf{y}^n)\right) + r_\gamma\left(\mathbf{y}^n, \Omega_\gamma(\mathbf{y}^n)\right),$$

so that $I_D(\mathbf{y}^n;\gamma) = D_D(\mathbf{y}^n;\gamma) - \log p_\gamma(\mathbf{y}^n|\hat{\boldsymbol{\theta}}_{\rm ML})$. We call $D_D(\cdot;\gamma)$ the *model cost*; it is the extra number of nits (nats) required to name the model used to transmit the data $\mathbf{y}^n$, the "model" being described by the uncertainty region $\Omega_\gamma(\mathbf{y}^n)$. In the case of MMLD the model cost is also the regret of the MMLD message length with respect to the "ideal" maximum likelihood codelength, though this is in general not the case for other MML approximations. In particular, comparing $D_D(\mathbf{y}^n;\gamma)$ to the MML87 model cost

$$D_{87}(\boldsymbol{\theta};\gamma) = -\log \pi_\gamma(\boldsymbol{\theta}) + \frac{1}{2}\log|\mathbf{J}_\gamma(\boldsymbol{\theta})| + \frac{k}{2}\left(\log \kappa_k + 1\right)$$

it is clear the fundamental difference between MML87 and MMLD is that $D_{87}(\boldsymbol{\theta};\gamma)$ depends on the chosen model $\boldsymbol{\theta}$ used to encode the data, while $D_D(\mathbf{y}^n;\gamma)$ depends on the data, and only offers a measure for complexity of a model structure γ. Thus, the MML87 codelength allows one to perform point estimation and model structure selection by minimising the sum of model cost for a particular model, say $\boldsymbol{\theta} \in \Theta_\gamma$, and the negative log-likelihood of the data using this model, i.e.,

$$\left\{\hat{\gamma}_{87}, \hat{\boldsymbol{\theta}}_{87}\right\} = \underset{\gamma\in\Gamma, \boldsymbol{\theta}\in\Theta_\gamma}{\arg\min}\left\{I(\gamma) + D_{87}(\boldsymbol{\theta};\gamma) - \log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})\right\}. \tag{6}$$

It would clearly be advantageous to have an analogue of (6) for a robust MMLD-like approximation.

3.2. MML08 Message Length

Examining (3) it is clear to see that $\mathbf{y}^n$ enters the round-off function (4) only through the negative log-likelihood function. Thus, following the arguments of Wallace and Freeman, we wish to find the *expected* increase in codelength due to quantisation of a model $\boldsymbol{\theta}^*$ to some region S. Rewrite $r_\gamma(\mathbf{y}^n, S)$ as

$$r_\gamma(\mathbf{y}^n, S) = -\frac{1}{q_\gamma(S)}\int_S \pi_\gamma(\boldsymbol{\theta})\log\frac{p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})}{p_\gamma(\mathbf{y}^n|\hat{\boldsymbol{\theta}}_{\rm ML})}d\boldsymbol{\theta}.$$

As in the Wallace–Freeman approximation [4], we can replace the dependency on a particular string $\mathbf{y}^n$ by a dependency on a particular model $\boldsymbol{\theta}^*$ by assuming that the data $\mathbf{y}^n \sim p_\gamma(\cdot|\boldsymbol{\theta}^*)$, and finding the expected inflation in codelength due to quantisation of $\boldsymbol{\theta}^*$. The average codelength for coding data $\mathbf{y}^n \sim p_\gamma(\cdot|\boldsymbol{\theta}^*)$ using model $\boldsymbol{\theta}$ is simply $\mathrm{E}_{\boldsymbol{\theta}^*}\left[\log 1/p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})\right]$, and this expression obtains a minimum when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ (the entropy). Thus, the expected excess codelength for coding data $\mathbf{y}^n$ coming from $p_\gamma(\cdot|\boldsymbol{\theta}^*)$ using model $\boldsymbol{\theta}$ in place of the optimal model $\boldsymbol{\theta}^*$ is

$$\Delta_\gamma(\boldsymbol{\theta}^*||\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}^*}\left[\log\left(\frac{p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}^*)}{p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})}\right)\right], \tag{7}$$

which is the well known Kullback–Leibler (KL) divergence [7] between a generating model $\boldsymbol{\theta}^*$ and approximating model $\boldsymbol{\theta}$. Note that the KL divergence in (7) is defined for n data points, which in the i.i.d. case is simply n times the KL divergence for a single datapoint.

Now the overall expected increase in codelength due to quantisation of the model $\boldsymbol{\theta}^*$ to the region S is given by

$$r_\gamma(\boldsymbol{\theta}^*, S) = \frac{1}{q_\gamma(S)}\int_S \pi_\gamma(\boldsymbol{\theta})\Delta_\gamma(\boldsymbol{\theta}^*||\boldsymbol{\theta})d\boldsymbol{\theta},$$

and we can find the quantisation cell that minimises the sum of the assertion plus the expected round off by solving

$$\Omega_\gamma(\boldsymbol{\theta}^*) = \underset{S\subseteq\Theta_\gamma}{\arg\min}\left\{\log 1/q_\gamma(S) + r_\gamma(\boldsymbol{\theta}^*, S)\right\}.$$

We call the set $\Omega_\gamma(\boldsymbol{\theta}^*)$ the *expected* uncertainty region for the model $\boldsymbol{\theta}^*$; in contrast to $\Omega_\gamma(\mathbf{y}^n)$, it depends only on the expected behaviour of the model $\boldsymbol{\theta}^*$. We now define the MML08 model cost for a model $\boldsymbol{\theta}^*$ by

$$D_{08}(\boldsymbol{\theta}^*;\gamma) = -\log q_\gamma\left(\Omega_n(\boldsymbol{\theta}^*)\right) + r_\gamma\left(\boldsymbol{\theta}^*, \Omega_\gamma(\boldsymbol{\theta}^*)\right). \tag{8}$$

Given that $D_{08}(\boldsymbol{\theta}^*;\gamma)$ depends only on the model, $\boldsymbol{\theta}^*$, and not on the data $\mathbf{y}^n$, we can compute a valid joint message length for any pair $(\boldsymbol{\theta}^*, \mathbf{y}^n) \in \Theta_\gamma \times \mathcal{Y}^n$; this is

$$I_{08}(\mathbf{y}^n, \boldsymbol{\theta}^*;\gamma) = D_{08}(\boldsymbol{\theta}^*;\gamma) - \log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}^*). \tag{9}$$

We call (9) the "MML08" message length approximation. Explicit point estimation, as well as model structure estimation, can be performed for a given $\mathbf{y}^n$ by comparing candidate models $\boldsymbol{\theta}^* \in \Theta_\gamma$, $\gamma \in \Gamma$, on their joint MML08 codelength, and choosing the model which yields the shortest message length, i.e.,

$$\left\{\hat{\gamma}_{08}, \hat{\boldsymbol{\theta}}_{08}\right\} = \underset{\gamma\in\Gamma, \boldsymbol{\theta}^*\in\Theta_\gamma}{\arg\min}\left\{I(\gamma) + I_{08}(\mathbf{y}^n, \boldsymbol{\theta}^*;\gamma)\right\}. \tag{10}$$

From (10) we see that the MML08 codelength balances the accuracy to which some particular model, $\boldsymbol{\theta}^* \in \Theta_\gamma$, is stated, against the evidence for that particular model that is present in the data. This is in contrast to the MMLD codelength, which implicitly quantises the maximum likelihood estimate. The MML08 message length may be split into assertion and detail components

$$I_{08}(\boldsymbol{\theta}^*;\gamma) = -\log q_\gamma\left(\Omega_\gamma(\boldsymbol{\theta}^*)\right),$$
$$I_{08}(\mathbf{y}^n|\boldsymbol{\theta}^*;\gamma) = -\log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}^*) + r_\gamma\left(\boldsymbol{\theta}^*, \Omega_\gamma(\boldsymbol{\theta}^*)\right).$$

The MML08 message length approximation generalizes the MMLD approximation, which can be recovered by setting $\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}_{\rm ML}$ and replacing the KL divergence with $\log p_\gamma(\mathbf{y}^n|\hat{\boldsymbol{\theta}}_{\rm ML})/p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})$, i.e., the empirical KL divergence.

Finally, we note that the MML08 model cost is very robust in the sense that $D_{08}(\boldsymbol{\theta}^*;\gamma) \geq 0$ for all $\boldsymbol{\theta}^* \in \Theta_\gamma$. In contrast, the MML87 model cost can be (nonsensically) negative if the conditions under which the MML87 approximation was derived are violated.

4. PROPERTIES OF MML08 CODELENGTHS

Assuming that the prior distribution and KL divergence are differentiable functions of $\boldsymbol{\theta}^*$, we have the following properties. The proofs are given in [5].

Property 1. The MML08 model cost (8) is invariant under differentiable, one-to-one transformations of the parameters $\boldsymbol{\theta}^*$, that is

$$I_{08}(\mathbf{y}^n, \boldsymbol{\theta}^*; \gamma) = I_{08}(\mathbf{y}^n, \boldsymbol{\phi}; \gamma),$$

where $\boldsymbol{\phi} = g(\boldsymbol{\theta}^*)$ are the transformed parameters, $g(\cdot)$ is a differentiable one-to-one function, and the prior distribution $\pi_\gamma(\boldsymbol{\theta}^*)$ is appropriately transformed.

This property has the important implication that inferences made by minimising the MML08 message length will be invariant to the choice of model parameterisation. This property is shared by the MML87 and SMML estimators, but not (in general) by Bayes estimators.

Property 2. The model cost (8) satisfies the "Boundary Rule" [1]; that is

$$\Omega_\gamma(\boldsymbol{\theta}^*) = \left\{ \boldsymbol{\theta} \in \Theta_\gamma : \Delta_\gamma(\boldsymbol{\theta}^* || \boldsymbol{\theta}) \leq \delta_\gamma(\boldsymbol{\theta}^*) \right\},$$

where $\delta_\gamma(\boldsymbol{\theta}^*)$ is the Kullback–Leibler divergence of any model on the boundary of $\Omega_\gamma(\boldsymbol{\theta}^*)$.

This property implies that the expected uncertainty region can be completely, and uniquely, defined by the value of the Kullback–Leibler divergence at the boundary of the region, $\delta_\gamma(\boldsymbol{\theta}^*)$. This property also suggests intriguing links with the normalised maximum likelihood code, and the concept of distinguishable distributions [2]; these links are interesting topics for future research.

5. LARGE SAMPLE BEHAVIOUR

The large sample behaviour of the MML08 approximation is now examined. Under the regularity conditions used in the derivation of the MML87 approximation [1], we have

$$\begin{aligned} \mathrm{vol}\left(\Omega_\gamma(\boldsymbol{\theta}^*)\right) \pi_\gamma(\boldsymbol{\theta}^*) &= q\left(\Omega_\gamma(\boldsymbol{\theta}^*)\right) + o_n(1), \\ \Delta_\gamma(\boldsymbol{\theta}^* || \boldsymbol{\theta}) &= \frac{1}{2} ||\boldsymbol{\theta} - \boldsymbol{\theta}^*||^2_{\mathbf{J}_\gamma(\boldsymbol{\theta}^*)} + o_n(1), \end{aligned}$$

where $||\mathbf{x}||^2_{\mathbf{A}} = \mathbf{x}'\mathbf{A}\mathbf{x}$. Ignoring terms of order $o_n(1)$, the KL divergence is a quadratic function of $\boldsymbol{\theta}$, and the uncertainty region will be a k-dimensional ellipse. Following similar arguments to those in [1], coupled with the rules for integration of polynomials over balls [8], the following assertion and detail lengths can be derived

$$\begin{aligned} I_{08}(\boldsymbol{\theta}; \gamma) \;=\; & -\log\left(\frac{\pi_\gamma(\boldsymbol{\theta})}{|\mathbf{J}_\gamma(\boldsymbol{\theta})|^{\frac{1}{2}}}\right) - \frac{k}{2}\log(\pi(k+2)) \\ & + \log\Gamma\left(\frac{k}{2}+1\right) + o_n(1), \quad (11) \end{aligned}$$

$$I_{08}(\mathbf{y}^n | \boldsymbol{\theta}; \gamma) \;=\; -\log p_\gamma(\mathbf{y}^n | \boldsymbol{\theta}) + \frac{k}{2} + o_n(1). \qquad (12)$$

Comparing (11–12) to (1–2), it is clear that for large n and sufficiently regular likelihood functions and prior distributions, the MML08 codelengths and MML87 codelengths differ only in their respective dimensionality constants. As (11–12) make use of elliptical uncertainty regions, which do not tessellate, the large sample MML08 codelength is actually slightly shorter than the MML87 codelength for $k > 1$. Interestingly, by assuming that the uncertainty region is congruent to an optimal quantising cell that tessellates the parameter space, the MML08 codelength can be used as a basis for a novel derivation of the MML87 approximation (as done in Chapter 2, [5]).

Under suitable regularity conditions, the large sample MML08 formulae (11–12) can be used to show that, asymptotically, as the sample size $n \to \infty$, with the number of parameters k fixed, the MML08 codelength is equivalent to the Bayesian information criterion [9]. The usual consistency properties of maximum likelihood parameter estimation, and Bayesian information criterion model structure selection, follow as a consequence.

6. ACKNOWLEDGEMENTS

The author would like to thank E. Makalic for comments and suggestions that improved this paper considerably.

7. REFERENCES

[1] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, Information Science and Statistics. Springer, first edition, 2005.

[2] J. J. Rissanen, *Information and Complexity in Statistical Modeling*, Information Science and Statistics. Springer, first edition, 2007.

[3] G. E. Farr and C. S. Wallace, "The complexity of strict minimum message length inference," *Computer Journal*, vol. 45, no. 3, pp. 285–292, 2002.

[4] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *Journal of the Royal Statistical Society (Series B)*, vol. 49, no. 3, pp. 240–252, 1987.

[5] D. F. Schmidt, *Minimum Message Length Inference of Autoregressive Moving Average Models*, Ph.D. thesis, Clayton School of Information Technology, Monash University, 2008.

[6] D. L. Dowe, "Foreword re: C. S. Wallace," *The Computer Journal*, vol. 51, no. 5, pp. 523–560, 2008.

[7] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, March 1951.

[8] G. B. Folland, "How to integrate a polynomial over a sphere," *The American Mathematical Monthly*, vol. 108, no. 5, pp. 446–448, 2001.

[9] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.