INTRODUCTION TO MINIMUM MESSAGE LENGTH INFERENCE

ENES MAKALIC AND DANIEL F. SCHMIDT

ABSTRACT. The aim of this manuscript is to introduce the Bayesian minimum message length principle of inductive inference to a general statistical audience that may not be familiar with information theoretic statistics. We describe two key minimum message length inference approaches and demonstrate how the principle can be used to develop a new Bayesian alternative to the frequentist t-test as well as new approaches to hypothesis testing for the correlation coefficient. Lastly, we compare the minimum message length approach to the closely related minimum description length principle and discuss similarities and differences between both approaches to inference.

1. MINIMUM MESSAGE LENGTH

Introduced in the late 1960s by Wallace and Boulton, the minimum message length (MML) principle [36, 41, 39, 35] is a framework for inductive inference rooted in information theory. The key idea underlying MML is that both parameter estimation and model selection problems can be interpreted as examples of data compression. It is well known that a random data string is not compressible. Conversely, if we have managed to compress a string of data we have (with high probability) learned something about the underlying structure of the data. Given data $\mathbf{y} \in \mathbb{R}^n$, the MML principle provides a method for computing the minimum length of a message that describes this data. In the MML approach, this message length is the single necessary inferential quantity. For the MML message to be decodable by a receiver, we require that the message encodes both a model for the data as well as the data itself. Formally, a MML message comprises two parts:

(1) the *assertion*: describes the structure of the model, including all model parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} \in \mathbb{R}^p$. Let $I(\boldsymbol{\theta})$ denote the codelength of the assertion.
(2) the *detail*: describes the data $\mathbf{y}$ using the model $p(\mathbf{y}|\boldsymbol{\theta})$ nominated in the assertion. Let $I(\mathbf{y}|\boldsymbol{\theta})$ denote the codelength of the detail.

The total length of the MML message, $I(\mathbf{y}, \boldsymbol{\theta})$, measured in units of information (for example, bits) is the sum of the lengths of the assertion and the detail:

$$(1) \qquad I(D, \boldsymbol{\theta}) = \underbrace{I(\boldsymbol{\theta})}_{\text{assertion}} + \underbrace{I(\mathbf{y}|\boldsymbol{\theta})}_{\text{detail}}.$$

The length of the assertion measures the complexity of the model, with longer assertions able to state more parameters with high accuracy or describe more complicated model structures. In contrast, a short assertion may encode the model parameters imprecisely and describe only simple models. The length of the detail tells us how well the model stated in the assertion is able to fit or describe the data. A complex model, that is one with a long assertion, will have lots of explanatory power and be able to encode more data strings using fewer bits compared to a simple model. MML seeks the model

$$(2) \qquad \hat{\boldsymbol{\theta}}(\mathbf{y}) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \left\{ I(\mathbf{y}, \boldsymbol{\theta}) \right\}$$

that minimises the length of the two-part message message. The key point is that minimising the two part message requires balancing the complexity of the model (assertion) with how well the model describes the data (detail). Ideally, we wish to find the simplest model that fits the observed data well enough; essentially, a formalisation of the famous razor of Occam. An advantage of MML is that the message length, measured in (say) bits, is a universal gauge that allows comparison across models with different model structures and numbers of parameters. As long as we can compute the MML codelengths of models, we can compare them. In this fashion, an MML practitioner is able to compare, for example, a linear regression model [32], to a finite mixture model [40] to a decision tree [43] via their codelengths for some observed data set.

The aim of this article is to introduce the MML principle of inductive inference to readers that may not be overly familiar with information theoretic statistics. We will examine two key MML codelength solutions to (2): the exact, or the so called Strict MML codelength (see Section 1.1), and a practical approximation to the exact codelength (see Section 1.2). We will demonstrate how to perform MML analysis of two common statistical problems: the two-sample t-test problem (see Section 2) and testing of the correlation coefficient (see Section 3). Lastly, we

compare MML to minimum description length (MDL), an alternative approach to inductive inference developed independently by Rissanen and colleagues [24, 25, 29, 28].

1.1. **Strict Minimum Message Length.** The exact solution to (2) is usually called Strict MML [37, 35], and may be viewed as a "gold standard' which more tractable formulae seek to approximate. To describe Strict MML we shall closely follow the notation in [35] (pp. 154–155). Let

$$(3) \qquad r(\mathbf{y}) = \int \pi(\boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) \, d\boldsymbol{\theta},$$

denote the marginal distribution of the n data points $\mathbf{y}$ given a prior distribution $\pi(\boldsymbol{\theta})$ over the model parameters $\boldsymbol{\theta} \in \mathbb{R}^p$. We note that MML is strictly Bayesian and requires a prior distribution for the model parameters in order to perform inference. Next we define the subset $\Theta^* = \{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots\} \subset \Theta$ as the set of parameters that can be used to encode the data $\mathbf{y}$, where each $\boldsymbol{\theta}_j^*$ may be encoded with different codelength. Recall that the assertion describes the model structure and all model parameters $\boldsymbol{\theta} \in \Theta$. To encode the assertion we require that Θ^* be a countable set as the assertion is finite and can encode only a countable subset of Θ. Clearly, in many inference problems the parameter set Θ is continuous and not countable and the question of how to quantize the parameter space Θ into a countable set Θ^* is a key issue in MML inference. Assuming we have somehow obtained the quantized parameter set Θ^*, the code for Θ^* implies a probability distribution over the set Θ^* since

$$p(\boldsymbol{\theta}_j^*) = q_j > 0, \quad \sum_{j:\boldsymbol{\theta}_j^* \in \Theta^*} q_j = 1, \quad (j = 1, 2, \ldots).$$

The assertion is completely determine by the set Θ^* and a distribution (code) over the set implying that a $\boldsymbol{\theta}_j^*$ should be optimally encoded with length $-\log q_j$.

The second part of the message, the detail, is the encoding of the data given the model $\hat{\boldsymbol{\theta}}(\equiv \boldsymbol{\theta}_j^*)$ specified in the assertion. Once we have chosen the model $\hat{\boldsymbol{\theta}}$, the data $\mathbf{y}$ can be encoded with length $-\log p(\mathbf{y}|\hat{\boldsymbol{\theta}})$. Let $t_j = \{\mathbf{y} \in \mathbb{R}^n : m(\mathbf{y}) = \boldsymbol{\theta}_j^*\}$ denote the set of data values $\mathbf{y}$ that will be encoded using some parameter $\boldsymbol{\theta}_j^* \in \Theta^*$ in the detail component of the message. Similarly, denote by $q_j = q(\boldsymbol{\theta}_j^*)$ the probability mass associated with parameter $\boldsymbol{\theta}_j^* \in \Theta^*$ used to construct the assertion codeword for $\boldsymbol{\theta}_j^*$. Strict MML seeks the mapping $m(\cdot)$ which minimizes the expected codelength of the message [35] (pp. 155)

$$(4) \qquad I_S = \sum_{\mathbf{y} \in \mathbb{R}^n} r(\mathbf{y}) \left[-\log q(m(\mathbf{y})) - \log p(\mathbf{y}|m(\mathbf{y})) \right],$$

where the expectation is taken with respect to the marginal distribution of the data. The first term in the expectation, $-\log q(m(\mathbf{y}))$, is the length of the codeword stating parameter $\boldsymbol{\theta}_j^* \in \Theta^*$ in the assertion, while the second term, $-\log p(\mathbf{y}|m(\mathbf{y}))$, denotes the codelength of a data set $\mathbf{y}$ encoded using parameter $\boldsymbol{\theta}_j^*$.

First, note that (4) minimizes the expected codelength over all possible data values because the map $m(\cdot)$ depends on unobserved data values that could be generated by $r(\mathbf{y})$, and not just on the particular data set that has been observed. Second, the expression for the expected codelength includes a sum over all possible data sets of size n as SMML considers the space of data to be countable. This is reasonable since we are interested in computing the length of a message that transmits the observed data and this message must be finite. The expected codelength can be written as

$$I_S = -\sum_{\boldsymbol{\theta}_j^* \in \Theta^*} \sum_{\mathbf{y}_i \in t_j} r(\mathbf{y}_i) \left(\log q(\boldsymbol{\theta}_j^*) + \log p(\mathbf{y}_i|\boldsymbol{\theta}_j^*) \right),$$

$$(5) \qquad = -\underbrace{\sum_{\boldsymbol{\theta}_j^* \in \Theta^*} q(\boldsymbol{\theta}_j) \log q(\boldsymbol{\theta}_j^*)}_{\text{assertion}} - \underbrace{\sum_{\boldsymbol{\theta}_j^* \in \Theta^*} \sum_{y_i \in t_j} r(\mathbf{y}_i) \log p(\mathbf{y}_i|\boldsymbol{\theta}_j^*)}_{\text{detail}},$$

where the coding probability of $\boldsymbol{\theta}_j^* \in \Theta^*$ is

$$(6) \qquad q(\boldsymbol{\theta}_j^*) = \sum_{\mathbf{y}_i \in t_j} r(\mathbf{y}_i),$$

which is the sum of the marginal probabilities of the data values that result in the estimate $\boldsymbol{\theta}_j^* \in \Theta^*$ being used in the detail. The first term in the expected Strict MML codelength is the *expected* length of the assertion, while the second term gives the expected length of the detail. Strict MML seeks to minimize the expected codelength since the map $m(\cdot)$ depends on all data values $\mathbf{y} \in \mathbb{R}^n$ and not just the observed data.

By seeking the optimal map $m(\cdot)$, Strict MML partitions the data space into disjoint regions t_j, such that all data sets that fall within a particular region t_j are encoded with the same parameter value $\boldsymbol{\theta}_j^* \in \Theta^*$. This partitioning

n	SMML PARTITION	Θ^*	I_S (bits)
1	{0..1}	{0.500}	1.000
2	{0..0, 1..2}	{0.000, 0.750}	1.667
3	{0..0, 1..3}	{0.000, 0.667}	2.085
4	{0..0, 1..3, 4..4}	{0.000, 0.500, 1.000}	2.454
5	{0..0, 1..4, 5..5}	{0.000, 0.500, 1.000}	2.704
6	{0..0, 1..5, 6..6}	{0.000, 0.500, 1.000}	2.962
7	{0..3, 4..7}	{0.214, 0.786}	3.165
8	{0..2, 3..7, 8..8}	{0.125, 0.625, 1.000}	3.337
9	{0..0, 1..5, 6..9}	{0.000, 0.333, 0.833}	3.491
10	{0..0, 1..4, 5..9, 10..10}	{0.000, 0.250, 0.700, 1.000}	3.647
11	{0..0, 1..5, 6..10, 11..11}	{0.000, 0.273, 0.727, 1.000}	3.762
12	{0..0, 1..6, 7..11, 12..12}	{0.000, 0.292, 0.750, 1.000}	3.887
13	{0..0, 1..6, 7..12, 13..13}	{0.000, 0.269, 0.731, 1.000}	3.998
14	{0..3, 4..10, 11..14}	{0.107, 0.500, 0.893}	4.107
15	{0..0, 1..5, 6..12, 13..15}	{0.000, 0.200, 0.600, 0.933}	4.204
16	{0..0, 1..5, 6..12, 13..16}	{0.000, 0.188, 0.563, 0.906}	4.289
17	{0..0, 1..6, 7..13, 14..17}	{0.000, 0.206, 0.588, 0.912}	4.372
18	{0..0, 1..5, 6..12, 13..17, 18..18}	{0.000, 0.167, 0.500, 0.833, 1.000}	4.457
19	{0..0, 1..5, 6..12, 13..18, 19..19}	{0.000, 0.158, 0.474, 0.816, 1.000}	4.531
20	{0..0, 1..6, 7..14, 15..19, 20..20}	{0.000, 0.175, 0.525, 0.850, 1.000}	4.601
21	{0..0, 1..6, 7..14, 15..20, 21..21}	{0.000, 0.167, 0.500, 0.833, 1.000}	4.665
22	{0..0, 1..6, 7..15, 16..21, 22..22}	{0.000, 0.159, 0.500, 0.841, 1.000}	4.737
23	{0..3, 4..11, 12..19, 20..23}	{0.065, 0.326, 0.674, 0.935}	4.801
24	{0..2, 3..9, 10..17, 18..23, 24..24}	{0.042, 0.250, 0.563, 0.854, 1.000}	4.863
25	{0..0, 1..6, 7..15, 16..22, 23..25}	{0.000, 0.140, 0.440, 0.760, 0.960}	4.920
26	{0..0, 1..6, 7..14, 15..22, 23..26}	{0.000, 0.135, 0.404, 0.712, 0.942}	4.974
27	{0..0, 1..6, 7..15, 16..23, 24..27}	{0.000, 0.130, 0.407, 0.722, 0.944}	5.025
28	{0..3, 4..12, 13..21, 22..27, 28..28}	{0.054, 0.286, 0.607, 0.875, 1.000}	5.080
29	{0..4, 5..13, 14..22, 23..28, 29..29}	{0.069, 0.310, 0.621, 0.879, 1.000}	5.129
30	{0..0, 1..5, 6..14, 15..23, 24..29, 30..30}	{0.000, 0.100, 0.333, 0.633, 0.883, 1.000}	5.176

TABLE 1. SMML partitions for a binomial distribution under a uniform prior on the probability of success. The countable set Θ^* denotes the parameters that would be used to code the data.

of the data space yields the quantized parameter set Θ^* where members of Θ^* are the feasible parameter estimates, with $m(\mathbf{y})$ being the point estimate associated with data string $\mathbf{y}$.

The countable set Θ^* that minimises (5) is not necessarily unique; however this is of no practical consequence as all optimal Θ^* result in an efficient encoding, and hence feasible parameter values [35] (pp. 161). The Strict MML codelength is invariant to data and model representation and represents the shortest two-part codelength for a message that conveys the data and a model for the data. Strict MML has very close links to fundamental concepts such as Turing machines and Kolmolgorov complexity [38]. We note that, while as have assumed the data space to be discrete in our exposition of Strict MML, this is not necessary and the Strict MML code for continuous data is a straightforward extension of the codelength (5); see [35], pp. 166–169.

Unfortunately, the computational complexity of partitioning the data space and obtaining Θ^* is NP complete in general and only practically computable for simple models [10]. To see this, we note that the Strict MML data space partitioning problem is an example of the well-known set partitioning problem (also known as the coalition structure generation problem in the artificial intelligence research community) provided the data space is countably finite [17, 23]. The set partitioning problem is known to be NP complete in general and cannot be solved by brute-force search as the total number of possible partitions, assuming a finite data space of size n, is the n-th Bell number B_n [5] and satisfies $\alpha n^n \leq B_n \leq n^n$ for some $\alpha > 0$ [31]. In some cases, constraints that impose structure and reduce the search space can be assumed. For example, if the data space is one-dimensional and ordered like in the example below, this is know as the ordered set partitioning problem and is identical to the shortest path problem that can be solved in $O(n^2)$ polynomial time.

Example: Consider data y representing the number of successes in n trials which follows the binomial distribution with unknown success probability $\theta \in \Theta = [0, 1]$; that is,

$$(7) \qquad p(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}.$$

Suppose that the success probability θ is assumed equally likely to take on any value in the parameter space Θ a priori; in other words, we assume the uninformative prior distribution $\pi(\theta) = 1$. The marginal distribution of the data is

$$r(y) = \int_0^1 p(y|\theta)\pi(\theta)d\theta = \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}d\theta = \frac{1}{n+1}.$$

To construct a Strict MML code for the binomial distribution we require the map $m(\cdot)$ that minimizes the expected codelength (5) over all possible data sets $y \in \mathbb{N} = \{0, 1, \ldots, n\}$, where the expectation is taken with respect to $r(y)$. This optimal codelength will be based on an optimal partitioning of the data space which implicitly yields the quantized parameter set Θ^*. The expected Strict MML codelength for the binomial distribution is

$$I_S = -\sum_{\theta_j^* \in \Theta^*} q(\theta_j^*)\log q(\theta_j^*) - \sum_{\theta_j^* \in \Theta^*}\sum_{y_i \in t_j}\left(\frac{1}{n+1}\right)\log\binom{n}{y_i}(\theta_j^*)^{y_i}(1-\theta_j^*)^{n-y_i}, \quad q(\theta_j^*) = \sum_{y_i \in t_j}\frac{1}{n+1},$$

and t_j is the set of data values y_i that will be encoded using $\theta_j^* \in \Theta^*$. For a given partition Θ^*, the contribution each estimate θ_j^* makes to the expected codelength is part of the detail only and does not influence the codelength of the assertion. To minimize the expected codelength for a particular θ_j^*, we therefore only need to minimize the second part of the message

$$-\sum_{y_i \in t_j} r(y_i)\log\binom{n}{y_i}(\theta_j^*)^{y_i}(1-\theta_j^*)^{n-y_i}.$$

Clearly, the $\hat{\theta}_j^*$ that minimizes the above is

$$\hat{\theta}_j^* = \frac{\sum_{y_i \in t_j} r(y_i)y_i}{n\sum_{y_i \in t_j} r(y_i)} = \frac{1}{n|t_j|}\sum_{y_i \in t_j} y_i, \quad j = 1, 2, \ldots,$$

where $|t_j|$ denotes the size of the set t_j; i.e., the number of data values that are encoded with parameter θ_j^*.

It remains to determine the optimal data space partitions t_j and hence the quantized set Θ^*. Before we look at the optimal Strict MML partitioning, we shall look at two examples with suboptimal codelengths. The simplest possible partitioning of the data space $\{0, 1, \ldots, n\}$ assumes that all data values are in the same partition, say, $t_1 = \{0, 1, \ldots, n\}$, where the size of the partition is $|t_1| = (n+1)$. This implies that the quantized parameter set $\Theta^* = \{\theta^*\}$ consists of a single member

$$\hat{\theta}^* = \frac{1}{n|t_1|}\sum_{y_i \in t_1} y_i = \frac{1}{2}.$$

which has probability $q(\theta^*) = 1$. The expected codelength of the assertion is then 0 bits and the expected codelength of the detail, and hence the total codelength, is

$$I_S = n\log 2 - \left(\frac{1}{n+1}\right)\sum_{i=0}^n \log\binom{n}{i}.$$

For example, for $n = 10$ and assuming a single partition that contains all possible data of size n, the Strict MML codelength for some observed data y is approximately 5.01 bits; here, the estimate of the success probability is $\hat{\theta} = 0.5$.

Alternatively, consider the more complex partitioning where the data space is divided into $(n+1)$ partitions with each datum $y_i \in \mathbb{N}$ belonging to its own partition. In this case, there are $(n+1)$ possible estimates and $\Theta^* = \{\theta_j^* : \theta_j^* = j/n, j = 0, 1, \ldots, n\}$. The probability of each estimate is

$$q_j = q(\theta_j^*) = \sum_{y_i \in t_j} r(y_i) = \frac{1}{n+1}, \quad j = 0, \ldots, n.$$

The expected Strict MML codelength is now

$$I_S = \log(n+1) - \frac{1}{n+1}\sum_{i=0}^n \log\binom{n}{i}\left(\frac{i}{n}\right)^i\left(1-\frac{i}{n}\right)^{n-i}.$$

For the $n = 10$ example, the total codelength of this partitioning is approximately 9.84 bits. This codelength is significantly worse than the single partition Strict MML code.

It is possible to compute the optimal Strict MML solution for the binomial distribution using the dynamic programming algorithm of Farr and Wallace [10]. The results of the algorithm are shown in Table 1 for all $1 \leq n \leq 30$; a MATLAB implementation of this algorithm can be downloaded from the Mathworks File Exchange website (ID: 80167). When $n = 10$, for example, Strict MML partitions the data space into four partitions $\{0, 1..4, 5..9, 10\}$

yielding the parameter set $\Theta^* = \{0.00, 0.25, 0.70, 1.00\}$. This implies that if we observe between 5 and 9 success counts, the estimated probability of success is 0.7. In contrast, if 0 or 10 success counts were observed, we would estimate the probability of success as 0 and 1, respectively. The Farr and Wallace algorithm is applicable to any one dimensional Strict MML problem with discrete data and is of polynomial complexity in terms of the sample size n.

An algorithm to obtain optimal Strict MML codelengths for one dimensional continuous data from exponential families with continuous sufficient statistics was developed by Dowty [9]. Unfortunately, there exists no general algorithm for computing optimal Stict MML partitions outside of the one dimensional (continuous or discrete) data setting. In the case of the k-nomial distribution with $k > 2$, the Strict MML data partitioning problem is an instance of the set partitioning (coalition structure generation) over graphs. While a dynamic programming algorithm exists to solve this problem, it runs in exponential time and is only practical for small sample sizes [20]. We note that many anytime exact algorithms and heuristic algorithms have also been proposed for this setting; see Rahwan et. al for a recent survey [23].

1.2. **A practical approximation to the Strict MML codelength.** Due to the high computational complexity of deriving the exact Strict MML codelength, Strict MML is mostly of interest from a theoretical standpoint. There exist several approximations to the Strict MML codelength with the MML87 approximation [41, 35] being the most popular. Under some regularity conditions [35]) (pp. 226), the MML87 codelength for data D is

$$(8) \qquad \mathcal{I}_{87}(\mathbf{y}, \boldsymbol{\theta}) = \underbrace{- \log \pi(\boldsymbol{\theta}) + \frac{1}{2} \log |J_{\boldsymbol{\theta}}(\boldsymbol{\theta})| + \frac{p}{2} \log \kappa_p}_{\text{assertion}} + \underbrace{\frac{p}{2} - \log p(\mathbf{y}|\boldsymbol{\theta})}_{\text{detail}}$$

where $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the prior distribution for the parameters $\boldsymbol{\theta}$, $|J_{\boldsymbol{\theta}}(\boldsymbol{\theta})|$ is the determinant of the expected Fisher information matrix, $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function of the model and κ_p is a quantization constant [7, 1]; for small p we have

$$(9) \qquad \kappa_1 = \frac{1}{12}, \quad \kappa_2 = \frac{5}{36\sqrt{3}}, \quad \kappa_3 = \frac{19}{192 \times 2^{1/3}},$$

while, for large p, κ_p is well-approximated by [35]:

$$(10) \qquad \frac{p}{2}(\log \kappa_p + 1) \approx -\frac{p}{2} \log 2\pi + \frac{1}{2} \log p\pi - \gamma,$$

where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant.

In contrast to SMML, which quantizes the parameter space by associating a single parameter estimate with a number of different data strings, the MML87 approximation achieves computational tractability by instead associating a set of models, deemed "indistinguishable" in an information sense, with the observed data string $\mathbf{y}$. This set of parameter estimates can be thought of as the range of possible parameter estimates that a plausible choice of $m(\cdot)$ would associate with the data string $\mathbf{y}$. The volume of this set of models, which is called the *uncertainty region* in MML parlance, explicitly determines the accuracy to which the continuous model parameters should be stated (i.e., the degree of quantization). We can re-write the codelength of the assertion in terms of the volume of the uncertainty region $w(\boldsymbol{\theta})$ as follows

$$(11) \qquad \underbrace{- \log \pi(\boldsymbol{\theta}) w(\boldsymbol{\theta})}_{\text{assertion}}, \quad w(\boldsymbol{\theta}) = \left(|J_{\boldsymbol{\theta}}(\boldsymbol{\theta})| \kappa_p^p \right)^{-1/2}.$$

It is clear that the volume of the uncertainty region depends on the variation of the negative log-likelihood function. If a small change in $\boldsymbol{\theta}$ results in a large change in the log-likelihood, the volume of the uncertainty region will be relatively small, implying that the parameters must be stated with higher precision. In contrast, if the negative log-likelihood is not particularly sensitive to small changes in $\boldsymbol{\theta}$, the volume of the uncertainty region will be fairly large. Under suitable regularity conditions, the volume of the uncertainty region is $O(n^{-p/2})$.

For many sufficiently well-behaved models, the MML87 codelength is virtually identical to the Strict MML codelength while being simpler to compute, requiring only the prior distribution for the model parameters and the determinant of the expected Fisher information matrix. Unlike the Strict MML codelength I_S (4) which must be found via minimization of an expectation, taken over all possible data with respect to the marginal distribution $r(\cdot)$, the MML87 codelength is computed using the observed data only. For large sample sizes $n \to \infty$, it is easy to show that the MML87 codelength is asymptotically equivalent to the well-known Bayesian information criterion (BIC) [33]

$$(12) \qquad \mathcal{I}_{87}(\mathbf{y}, \boldsymbol{\theta}) = - \log p(\mathbf{y}|\boldsymbol{\theta}) + \frac{p}{2} + O(1),$$

where the $O(1)$ term depends on the prior distribution, the Fisher information and the number of parameters p. In fact, as the MML87 codelength $\mathcal{I}_{87}(\mathbf{y}, \boldsymbol{\theta})$ can be interpreted as the negative logarithm of the posterior probability mass attached to a dataset $\mathbf{y}$ and model $\boldsymbol{\theta}$, the difference in message lengths admits the interpretation as the logarithm of the posterior-odds of two models, e.g.,

$$\mathcal{I}_{87}(\mathbf{y}, \boldsymbol{\theta}_0) - \mathcal{I}_{87}(\mathbf{y}, \boldsymbol{\theta}_1)$$

can be interpretated as the posterior log-odds in favour of model $\boldsymbol{\theta}_1$ against $\boldsymbol{\theta}_0$. The MML87 codelength results in estimates that are invariant under (smooth) one-to-one reparametarization, just like the maximum likelihood estimate. MML87 has been applied to a wide range of statistical models including decision tress [43], factor analysis [42] and mixture models [40].

Example: We will compute the MML87 codelength for the binomial distribution and compare it to the Strict MML code [35] (pp. 246). Let $y \in [0, n]$ be the count of successes which follows a binomial distribution with probability density function (7) where $0 < \theta < 1$ is the probability of observing a success. Suppose we assume a uniform prior distribution for the success probability $\pi(\theta) = 1$. The determinant of the expected Fisher information matrix for a binomial distribution is $J(\theta) = n/(\theta(1 - \theta))$. The volume of the uncertainty region is

$$w(\theta) = \left(\frac{12(1 - \theta)\theta}{n} \right)^{1/2}$$

which gets smaller as $n \to \infty$, or as the success probability θ approaches $\theta \to 0$ or $\theta \to 1$. Substituting into the MML87 codelength (8) we obtain

$$\mathcal{I}_{87}(y, \theta) = \frac{1}{2} \log \left(\frac{n}{\theta(1 - \theta)} \right) + \frac{1}{2} \log \frac{1}{12} - \log \binom{n}{y} \theta^y (1 - \theta)^{n-y} + \frac{1}{2}$$

$$(13) \qquad\qquad = -\left(y + \frac{1}{2} \right) \log \theta - \left(n - y + \frac{1}{2} \right) \log(1 - \theta) + \frac{1}{2} \left(1 + \log \frac{n}{12 \binom{n}{y}^2} \right)$$

This codelength is minimized by choosing θ to be the MML87 estimate

$$(14) \qquad\qquad \hat{\theta}(y) = \frac{y + 1/2}{n + 1},$$

which results in the optimal MML87 codelength for data y being

$$\mathcal{I}_{87}(y) = -\left(y + \frac{1}{2} \right) \log \hat{\theta}(y) - \left(n - y + \frac{1}{2} \right) \log(1 - \hat{\theta}(y)) + \frac{1}{2} \left(1 + \log \frac{n}{12 \binom{n}{y}^2} \right).$$

As an example, if $n = 10$ and $y = 3$, the MML87 estimate of the success probability is $\hat{\theta}(y) = 7/22 \approx 0.32$, the volume of the uncertainty region is $w(\theta) \approx 0.51$ and the MML87 codelength is approximately 3.61 bits. In contrast, the SMML estimate of the success probability is $\hat{\theta}(y) = 0.25$ resulting in a codelength of 3.647 bits. We can empirically show that the expected MML87 codelength with respect to the marginal distribution $r(y) = 1/(n + 1)$ falls within 0.1 bits of the expected Strict MML codelength for all $n \geq 5$, while being significantly simpler to compute.

2. Minimum message length t-test

Perhaps the most popular hypothesis test in practice is the frequentist two-sample t-test for comparison of two means. As an example of the t-test popularity, a search for the term 'paired t-tests' returns approximately 12,000 articles from 1990 onwards in PubMed. Similarly, articles published in peer reviewed psychology journals feature, on average, more than three t-tests per article [45]. Formally, the t-test assumes that we observe normally distributed data from two, possibly different, populations:

$$(15) \qquad\qquad Y_{1i} \sim \mathcal{N} \left(\mu + \frac{\sigma\delta}{2}, \sigma^2 \right), \quad Y_{2j} \sim \mathcal{N} \left(\mu - \frac{\sigma\delta}{2}, \sigma^2 \right),$$

for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. Here, the unknown parameter $\mu \in \mathbb{R}$ is the overall (grand) mean, $\sigma > 0$ is a standard deviation that is common to both groups, $\delta > 0$ is the standardised effect size and (n_1, n_2) are the sample sizes of the two groups. This formulation of the t-test in terms of the overall mean and standardised effect

size is due to [11]. The usual null hypothesis says that the mean is the same in both groups while the alternative hypothesis specifies a different mean for each group; formally, we have

(16) $$\mathcal{H}_0 : \delta = 0, \quad \mathcal{H}_1 : \delta \neq 0.$$

Let $\boldsymbol{\theta}_0 = \{\mu, \sigma\} \in (\mathbb{R} \times \mathbb{R}_+) \equiv \Theta_0$ and $\boldsymbol{\theta}_1 = \{\mu, \sigma, \delta\} \in (\mathbb{R}^2 \times \mathbb{R}_+) \equiv \Theta_1$ denote the parameter set under the two models. Given data $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2\}$ where $\mathbf{y}_1 = (y_{1,1}, \ldots, y_{n_1,1})'$ and $\mathbf{y}_2 = (y_{2,1}, \ldots, y_{n_2,1})'$, the task is estimate the effect size $\delta > 0$ and thus determine whether the population means of the two groups are equal ($\delta = 0$).

Recently, there has been a great deal of interest in Bayesian alternatives to the frequentist t-test [11, 30, 44, 12, 15]. A key quantity in Bayesian analysis is the Bayes factor

(17) $$\underbrace{\frac{P(\mathcal{H}_1|\mathbf{y})}{P(\mathcal{H}_0|\mathbf{y})}}_{\text{Posterior odds}} = \underbrace{\frac{p(\mathbf{y}|\mathcal{H}_1)}{p(\mathbf{y}|\mathcal{H}_0)}}_{\text{Bayes factor}} \underbrace{\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_0)}}_{\text{Prior odds}},$$

which is the ratio of the marginal likelihoods of the data obtained by integrating the parameters out with respect to their prior distribution; that is,

(18) $$p(\mathbf{y}|\mathcal{H}_i) = \int p(\mathbf{y}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}, \quad i = 0, 1,$$

where $\pi_i(\boldsymbol{\theta}_i)$ is the prior distribution of the parameters under model i. The Bayes factor, henceforth denoted by BF_{10}, measures the amount of evidence in support of hypothesis $\mathcal{H}_1$ over hypothesis $\mathcal{H}_0$, where $\text{BF}_{10} > 3$ indicates substantial evidence in favour of $\mathcal{H}_1$ [14]. Given a proper prior $\pi(\delta)$ on the effect size δ, Gronau et al. [12], generalizing the work of [11, 30], showed that the Bayes factor for the t-test problem can be expressed as

(19) $$\text{BF}_{10} = \frac{\int T_\nu(t|\sqrt{n_\delta}\delta)\pi(\delta)\, d\delta}{T_\nu(t)}$$

where t is the observed t-statistic

(20) $$t = \sqrt{n_\delta}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)/s_p, \quad \nu s_p = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2,$$

and s_j $(j = 1, 2)$ are the usual unbiased sample variance estimates. Here, $\nu = n_1 + n_2 - 2$ denotes the degrees of freedom, $n_\delta = (1/n_1 + 1/n_2)^{-1}$ is the effective sample size, and $T_\nu(\cdot|a)$ is the non-central Student t distribution with degrees of freedom ν and non-centrality parameter a. For most prior distributions $\pi(\delta)$, the integral in the Bayes factor must be computed by numerical integration.

In order to perform the two-sample t test within the MML framework, we require the codelengths of the data under the null distribution ($\delta = 0$) and under the alternative hypothesis ($\delta \neq 0$). Under the null model, the negative log-likelihood of the data $\mathbf{y}$ is

$$-\log \ell_0(\boldsymbol{\theta}_0) = \frac{n_1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n_1} (y_{1i} - \mu)^2 + \frac{n_2}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{j=1}^{n_2} (y_{2j} - \mu)^2,$$

$$= \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2,$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ denotes the two-samples stacked into a single data vector of length $n = (n_1 + n_2)$. Let $\pi_0(\boldsymbol{\theta}_0)$ denote the prior distributions for the parameters under the null hypothesis. We use the usual right Haar prior for the mean and standard deviation:

(21) $$\pi_0(\boldsymbol{\theta}_0) = (\Omega\sigma)^{-1}$$

defined over some suitable parameter range Ω implying a location and scale invariant distribution of the mean and standard deviation, respectively. The determinant of the expected Fisher information matrix for the normal distribution is $|J(\boldsymbol{\theta}_0)| = 2n^2/\sigma^4$ leading to the MML87 codelength:

$$\mathcal{I}_0(\mathbf{y}, \boldsymbol{\theta}_0) = -\log \ell_0(\boldsymbol{\theta}_0) + \frac{1}{2} \log |J(\boldsymbol{\theta}_0)| - \log \pi_0(\boldsymbol{\theta}_0) + \log \kappa_2 + 1$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 + (n-1)\log(\sigma) + \frac{1}{2} \log\left(2^{n+1}\pi^n (ne\kappa_2\Omega)^2\right)$$
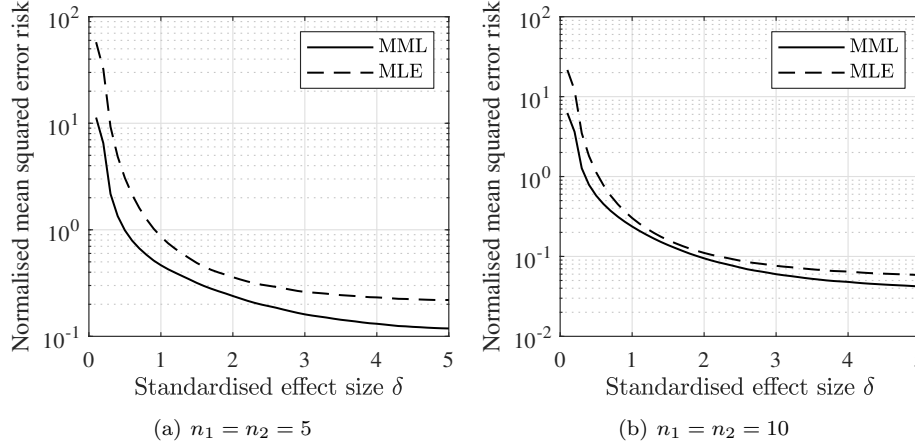
(a) $n_1 = n_2 = 5$ (b) $n_1 = n_2 = 10$

FIGURE 1. Normalised mean square error for the maximum likelihood and MML estimates of δ with $n_1 = n_2 = 5$ and $n_1 = n_2 = 10$ data samples averaged over 10^5 simulations. In all cases, the mean and standard deviation of the data generating model were $\mu = 0, \sigma = 1$.

Recall that we can interpret the difference in MML codelengths as the log posterior probability in favour of the model with the shorter codelength.

A brief simulation was performed to compare the behaviour of the MML87 estimate of the standardised effect size δ to the ML estimate. Data with sample sizes $(n_1, n_2) \in \{5, 10, 20, 50\}$ was generated from the model ($\mu = 0, \sigma = 1$) with $0.1 \leq \delta \leq 5$ and both ML and MML87 were asked to nominate an estimate of δ. The experiment was repeated for 10^5 iterations and the estimates were compared using the normalized mean squared error metric

$$\text{(29)} \qquad \text{NMSE} = \frac{(\delta - \hat{\delta}(\mathbf{y}))^2}{\delta},$$

averaged over all 10^5 iterations for each experiment. Results for sample sizes $n_1 = n_2 = 5$ and $n_1 = n_2 = 10$ are shown in Figure 1. Behaviour of the ML and MML87 estimates for moderate to large sample sizes was indistinguishable, as predicted by asymptotic analysis. When the sample size was small, ML tends to overestimate δ and underestimate σ. This is in contrast to MML87 which tends to underestimate δ and is more conservative.

In terms of model (hypothesis) selection, unlike the frequentist methodology, neither the Bayes factor nor MML approach are designed to explicitly control type I error. Empirically, we observe that both the Bayes factor derived by Gronau et al [12] and the MML87 t-test derived in Section 2 control type I error rate at approximately 0.10 for all $0.1 < \delta < 2$, if the alternative model is chosen whenever $\mathcal{I}_1(y) < \mathcal{I}_0(y)$, or equivalently, whenever the Bayes factor $\text{BF}_{10} > 1$. To control the type I error rate at 0.05, we would need to select the alternative model for all $\text{BF}_{10} > 1.87$ or whenever $(\mathcal{I}_1(y) + \log(1.65)) < \mathcal{I}_0(y)$. Under this setup, all three methods were virtually indistinguishable in terms of type I error rate for all $0.1 < \delta < 2$.

3. MINIMUM MESSAGE LENGTH TESTING OF THE CORRELATION COEFFICIENT

Suppose that a pair of random variables (Y_1, Y_2) follows a bivariate normal distribution with means (μ_1, μ_2), variances (σ_1^2, σ_2^2) and correlation coefficient $\rho \in (0, 1)$. We have n observations $\mathbf{y} = \{(y_{1,i}, y_{2,i})\}$ $(i = 1, \ldots, n)$ and wish to test the hypothesis

$$H_0 : \rho = \rho_0 \quad \text{versus} \quad H_1 : \rho \neq \rho_0$$

where $\rho_0 \in (0, 1)$ is a fixed, user-specified value. Let $\boldsymbol{\theta}_0 = \{\mu_1, \mu_2, \sigma_1, \sigma_2\} \in \mathbb{R}^2 \times \mathbb{R}_+^2 \equiv \Theta_0$ and $\boldsymbol{\theta}_1 = \{\mu_1, \mu_2, \sigma_1, \sigma_2, \rho\} \in \mathbb{R}^2 \times \mathbb{R}_+^2 \times (0, 1) \equiv \Theta_1$ denote the parameter sets under the null and alternative hypothesis models, respectively. A large number of frequentist as well as Bayesian procedures exist for this hypothesis testing problem. A good summary of the recent literature is in [22], which also introduces a new Bayesian procedure for testing the correlation coefficient under divergence-based priors [4].

Under the null model, the negative log-likelihood of the data $\mathbf{y}$ is

$$\text{(30)} \qquad -\log \ell_0(\boldsymbol{\theta}_0) = n \log(2\pi) + n \log(\sigma_1 \sigma_2) + \frac{n}{2} \log(1 - \rho_0^2) + \frac{1}{2(1 - \rho_0^2)} \sum_{i=1}^{n} Q_i(\mathbf{y}, \boldsymbol{\theta}_0)$$

where

$$Q_i(\mathbf{y}, \boldsymbol{\theta}) = \frac{(y_{1,i} - \mu_1)^2}{\sigma_1^2} - \frac{2\rho\,(y_{1,i} - \mu_1)\,(y_{2,i} - \mu_2)}{\sigma_1 \sigma_2} + \frac{(y_{2,i} - \mu_2)^2}{\sigma_2^2}. \tag{31}$$

Let

$$\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{j,i}, \quad s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{j,i} - \bar{y}_j)^2, \quad r = \frac{1}{n s_1 s_2}\sum_{i=1}^{n}(y_{1,i} - \bar{y}_1)(y_{2,i} - \bar{y}_2), \tag{32}$$

for j = 1,2 denote the usual sufficient statistics. Maximum likelihood estimates of the model parameters are

$$\hat{\mu}_j(\mathbf{y})_{\mathrm{ML}} = \bar{y}_j, \quad \hat{\sigma}_j^2(\mathbf{y})_{\mathrm{ML}} = s_j^2\left(\frac{1 - \rho_0\,r}{1 - \rho_0^2}\right), \quad (j = 1, 2). \tag{33}$$

For the MML codelength, we again use the right Haar prior for the means and standard deviations

$$\pi_0(\boldsymbol{\theta}_0) = (\Omega_0 \sigma_1 \sigma_2)^{-1} \tag{34}$$

where Ω_0 is a normalisation constant. The determinant of the expected Fisher information matrix for a bivariate normal distribution with a fixed (known) correlation coefficient ρ_0 is

$$|J(\boldsymbol{\theta}_0)| = \frac{4n^4}{(1 - \rho_0^2)^2 \sigma_1^4 \sigma_2^4} \tag{35}$$

and the MML87 codelength for the data $\mathbf{y}$ is

$$\begin{aligned}
\mathcal{I}_0(\mathbf{y}, \boldsymbol{\theta}_0) &= -\log \ell_0(\boldsymbol{\theta}_0) + \frac{1}{2}\log|J(\boldsymbol{\theta}_0)| - \log \pi_0(\boldsymbol{\theta}_0) + 2(1 + \log \kappa_4) \\
&= (n-1)\log(\sigma_1 \sigma_2) + \frac{n-2}{2}\log(1 - \rho_0^2) + \frac{1}{2\,(1 - \rho_0^2)}\sum_{i=1}^{n} Q_i(\mathbf{y}, \boldsymbol{\theta}_0) \\
&\quad + \log\left(\Omega_0 \pi^n 2^{n+1}(\kappa\,n)^2\right) + 2
\end{aligned}$$

where $\kappa_4 \approx 0.076603$. MML estimates that minimise the codelength under the null hypothesis are

$$\hat{\mu}_j(\mathbf{y}) = \bar{y}_j, \quad \hat{\sigma}_j^2(\mathbf{y}) = s_j^2\left(\frac{n(1 - \rho_0\,r)}{(n-1)\,(1 - \rho_0^2)}\right), \quad (j = 1, 2). \tag{36}$$

We observe that the MML estimates of the mean parameters are the same as the usual maximum likelihood estimates. However, MML estimates of the variances include the extra term $n/(n-1)$. If the null hypothesis corresponds to the no correlation model ($\rho_0 = 0$), MML estimates of the variances are unbiased unlike the corresponding maximum likelihood estimates.

Under the alternative hypothesis, the correlation coefficient ρ is unknown and must be estimated from the data along with the means and variances. The negative log-likelihood of the data $\mathbf{y}$ is now

$$-\log \ell_1(\boldsymbol{\theta}_1) = n\log(2\pi) + n\log(\sigma_1 \sigma_2) + \frac{n}{2}\log(1 - \rho^2) + \frac{1}{2\,(1 - \rho^2)}\sum_{i=1}^{n} Q_i(\mathbf{y}, \boldsymbol{\theta}_1) \tag{37}$$

Maximum likelihood estimates of the parameters are

$$\hat{\mu}_j(\mathbf{y})_{\mathrm{ML}} = \bar{y}_j, \quad \hat{\sigma}_j^2(\mathbf{y})_{\mathrm{ML}} = s_j^2, \quad \hat{\rho}(\mathbf{y})_{\mathrm{ML}} = r, \quad (j = 1, 2). \tag{38}$$

We use the same right Haar prior for the means and standard deviations and opt for the uniform prior distribution over the correlation coefficient leading to

$$\pi_1(\boldsymbol{\theta}_1) = (\Omega_1 \sigma_1 \sigma_2)^{-1} \tag{39}$$

where Ω_1 is a normalisation constant. The determinant of the expected Fisher information matrix for a bivariate normal distribution with an unknown correlation coefficient is

$$|J(\boldsymbol{\theta}_1)| = \frac{4n^5}{(1 - \rho^2)^4 \sigma_1^4 \sigma_2^4}. \tag{40}$$

MML codelength of the alternative hypothesis is

$$\mathcal{I}_1(\mathbf{y}, \boldsymbol{\theta}_1) = -\log \ell_1(\boldsymbol{\theta}_1) + \frac{1}{2}\log |J(\boldsymbol{\theta}_1)| - \log \pi_1(\boldsymbol{\theta}_1) + \frac{5}{2}(1 + \log \kappa_5)$$

$$= (n-1)\log(\sigma_1 \sigma_2) + \frac{n-4}{2}\log(1-\rho^2) + \frac{1}{2(1-\rho^2)}\sum_{i=1}^{n} Q_i(\mathbf{y}, \boldsymbol{\theta}_1)$$

$$+ \log\left(\Omega_1 \pi^n 2^{n+1}(n\,\kappa_5)^{5/2}\right) + \frac{5}{2}$$

where $\kappa_5 \approx 0.075625$. MML estimates that minimise the codelength are

$$\hat{\mu}_j(\mathbf{y}) = \bar{y}_j,$$

$$\hat{\sigma}_j^2(\mathbf{y}) = s_j^2\left(\frac{n(n+2+\sqrt{(n+2)^2 - 12r^2(n-1)})}{2(n+2)(n-1)}\right) = s_j^2\left(\frac{n(n-3\hat{\rho}r+2)}{(n-1)(n+2)}\right)$$

$$\hat{\rho}(\mathbf{y}) = \frac{n+2-\sqrt{(n+2)^2 - 12r^2(n-1)}}{6r} = \frac{(n+2)}{3r}\left(1 - \left(1 - \frac{1}{n}\right)\left(\frac{\hat{\sigma}_j}{s_j}\right)^2\right).$$

$$= r\left(\frac{n-1}{n+2}\right) + O\left(n^{-1}\right)$$

for $j = 1, 2$. It is easy to show that the MML estimate of the correlation parameter is always smaller in magnitude compared to the sample correlation coefficient; ie, $|\hat{\rho}(\mathbf{y})| < |r|$. Figure 2 shows the mean squared error (MSE) risk for $n \in \{20, 50\}$ for the MML, maximum likelihood and the unique minimum variance unbiased estimator [21]

$$\hat{\rho}(\mathbf{y})_{\text{Unbiased}} = r\,_2F_1\left(\frac{1}{2}, \frac{1}{2}; \frac{n-1}{2}; 1 - r^2\right)$$

of the correlation parameter ρ; here, $_2F_1(\cdot)$ is Gaussian hypergeometric function. When $n = 20$, the MML estimate dominates the maximum likelihood estimate in terms of MSE risk for all $|\rho| < 0.64$. As $|\rho| \to 1$, all three estimates are virtually indistinguishable. When $n = 50$, the MML estimate again dominates the maximum likelihood estimate for all $|\rho| < 0.6$.

(a) $n = 20$ (b) $n = 50$
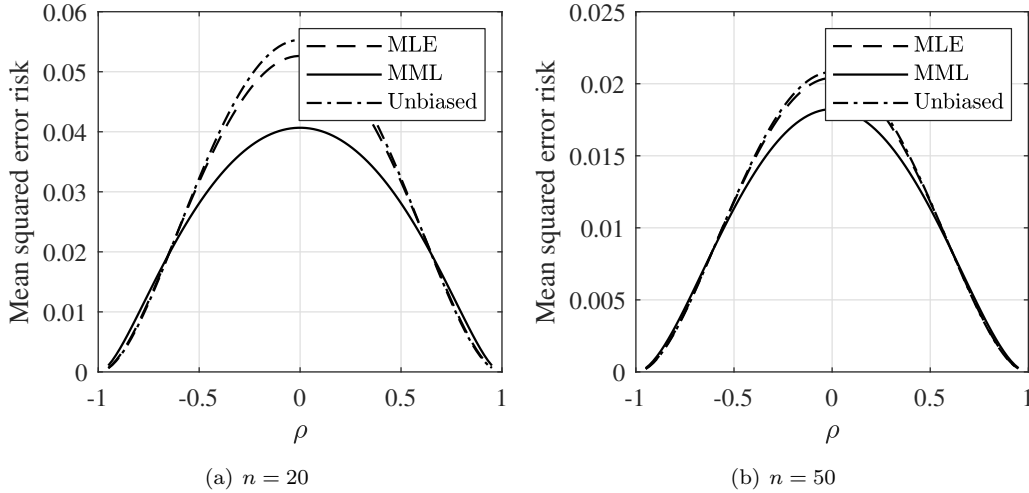
FIGURE 2. Expected mean squared error for the MML, maximum likelihood and the unbiased minimum variance estimators of the correlation parameter ρ.

4. DISCUSSION

An advantage of MML over alternative approaches is that it allows conceptually straightforward comparison of models with different structures and parameters as long as we can compute their codelengths. The codelength is measured in the unit of information (e.g., bits) and is therefore a universal yardstick for model comparison. It is

ρ_0	ρ	n	Relative frequency of rejecting $H_0 : \rho = \rho_0$							KL Divergence	
			DB^S	DB^M	Fisher Z	HP	Mud	Rud	MML	MLE	MML
-0.30	-0.75	15	0.730	**0.812**	0.704	0.730	0.702	0.679	0.570	**0.408**	0.414
		30	0.921	**0.950**	0.936	0.938	0.936	0.932	0.890	**0.129**	0.145
	-0.30	15	0.073	0.124	0.060	0.068	0.059	0.058	**0.050**	0.262	**0.230**
		30	0.043	0.081	0.046	0.050	0.046	0.046	**0.035**	0.132	**0.124**
	0.00	15	0.236	**0.329**	0.184	0.206	0.178	0.193	0.245	0.229	**0.190**
		30	0.354	**0.465**	0.371	0.382	0.371	0.382	0.358	0.092	**0.083**
	0.50	15	0.908	**0.934**	0.883	0.893	0.881	0.887	0.904	0.278	**0.223**
		30	0.995	**0.998**	0.995	0.996	0.995	0.995	0.994	0.105	**0.097**
0.00	-0.75	15	0.969	**0.988**	0.953	0.960	0.950	0.950	0.938	0.298	**0.252**
		30	0.998	0.998	0.998	0.998	0.998	0.998	**0.999**	0.106	**0.099**
	-0.30	15	0.262	**0.362**	0.207	0.224	0.204	0.204	0.198	0.277	**0.237**
		30	0.352	**0.476**	0.356	0.368	0.356	0.356	0.322	0.126	**0.118**
	0.00	15	0.074	0.130	0.055	0.060	0.055	0.055	**0.051**	0.217	**0.189**
		30	0.050	0.106	0.053	0.057	0.053	0.053	**0.037**	0.087	**0.080**
	0.50	15	0.613	**0.700**	0.559	0.580	0.554	0.548	0.508	0.330	**0.277**
		30	0.809	**0.893**	0.820	0.057	0.053	0.053	0.791	0.126	**0.120**
0.70	-0.75	15	**1.000**	1.000	1.000	1.000	1.000	1.000	1.000	0.273	**0.232**
		30	**1.000**	1.000	1.000	1.000	1.000	1.000	1.000	0.106	**0.098**
	-0.30	15	0.984	0.987	0.985	0.986	0.985	0.986	**0.992**	0.264	**0.221**
		30	**1.000**	1.000	1.000	1.000	1.000	1.000	1.000	0.105	**0.096**
	0.00	15	0.834	0.875	0.850	0.865	0.849	0.867	**0.910**	0.242	**0.207**
		30	0.992	0.998	0.998	0.998	0.998	**0.999**	0.993	0.104	**0.095**
	0.50	15	0.131	0.182	0.150	0.167	0.149	0.173	**0.258**	0.328	**0.287**
		30	0.275	0.397	0.403	0.412	0.401	**0.417**	0.354	**0.176**	0.178

TABLE 2. Relative frequency of rejecting the null hypothesis and Kullback–Leibler (KL) divergence from the data generating model for maximum likelihood (MLE) and minimum message length (MML) estimates.

then easy to extend the MML t-test to, for example, handle the case of unequal variances in the two groups, a version of the so-called Behrens–Fisher problem. The setup is now

$$\text{(41)} \qquad Y_{1i} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right), \quad Y_{2j} \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right),$$

for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$ and the task is, given data $\mathbf{y}$, to determine whether the two means are equal $\mu_1 = \mu_2$. MML87 codelengths for this setup were derived in [18] and can be compared to the codelengths of the null and alternative hypotheses under the equal variance assumption derived in Section 2 to discriminate between the four candidate models. Of course, we can also include other models in the comparison, as long as we can compute the corresponding codelengths.

4.1. **Minimum description length.** Minimum description length (MDL) [24, 25, 29, 28, 13, 6] is a closely related approach to inductive inference that was developed around the same time as MML by Rissanen and colleagues and has gone through various refinements. MDL seeks a model class $\mathcal{M}$ that results in the shortest codelength of the data (ie, a model that best compresses the data), where a model class is defined as a set of parametric distributions indexed by parameter $\boldsymbol{\theta}$:

$$\mathcal{M} = \{p(\cdot|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}.$$

A popular version of MDL is the general normalized maximum likelihood (NML) code whose codelength is given by

$$\text{(42)} \qquad -\log p_{\mathrm{NML}}(\mathbf{y}|v, \mathcal{M}) = \underbrace{-\log\left[\max_{\boldsymbol{\theta}\in\Theta} p(\mathbf{y}|\boldsymbol{\theta}(\mathbf{y}), \mathcal{M})\, v(\boldsymbol{\theta})\right]}_{\text{model fit}} + \underbrace{\log\left[\sum_{\mathbf{x}} \max_{\boldsymbol{\theta}\in\Theta} p(\mathbf{x}|\boldsymbol{\theta}(\mathrm{x}), \mathcal{M})\, v(\boldsymbol{\theta})\right]}_{\text{parametric complexity}}$$

where $v : \Theta \to [0, \infty)$ is a user-specified function, referred to as the *luckiness* function in MDL terminology, and the sum (integral) is over the entire data space in the case of discrete (continuous) data. This is the general NML codelength for data $\mathbf{y}$ with respect to some model class $\mathcal{M}$ parameterised by models $\boldsymbol{\theta} \in \Theta \in \mathcal{M}$. We observe that the NML codelength is a sum of two terms:

   (1) the (penalized) negative log-likelihood of the data evaluated at the maximum, and
   (2) the so-called parametric complexity of the model class $\mathcal{M}$ which measures how well models $\boldsymbol{\theta} \in \mathcal{M}$ approximate random data sequences.

The first term in the NML codelength represents how well the model $\hat{\theta} \in \mathcal{M}$ fits the observed data with smaller codelengths indicating better data fit. In contrast, a low (high) parametric complexity implies that a few (many) data sequences can be modelled sufficiently well by models in class $\mathcal{M}$. The parametric complexity denotes the logarithm of the number of distinguishable distributions in the model class [2] and is similar in spirit to the assertion in the MML code, accounting for the complexity of the model class from which the maximum likelihood estimate is derived. A detailed discussion of the NML distribution and the corresponding optimality properties is available in [26, 27, 13].

Originally, Risannen advocated the use of the NML distribution for $v(\theta) = 1$ only. Unfortunately, the parametric complexity in this case is easily shown to be infinite for many model classes rendering a straightforward application of the NML codelength impossible. Subsequently, Rissanen and colleagues suggested other forms of the luckiness function as well as variations of the NML code such as the restricted approximate normalized maximum likelihood (ANML), the two-part ANML or the objective Bayesian code, among others [8, 13].

Example: Assuming $v(\theta) = 1$, the NML codelength for the binomial distribution is

$$(43) \qquad -\log p_{\mathrm{NML}}(y|\mathcal{M}) = -\log \binom{n}{y} \hat{\theta}^y (1-\hat{\theta})^{n-y} + \log \mathcal{C}(n)$$

where

$$(44) \qquad \mathcal{C}(n) = \sum_{x=0}^{n} \binom{n}{x} \left(\frac{x}{n}\right)^x \left(\frac{n-x}{n}\right)^{n-x}$$

and $\hat{\theta}(y) = (y/n)$ is the maximum likelihood estimate of the success probability. An efficient linear-time algorithm to compute the parametric complexity of a general multinomial distribution is available in [16]. As an example, if $n = 10$ and $y = 3$, the maximum likelihood estimate is $\hat{\theta}(y) = 0.3$, the parametric complexity is $\log \mathcal{C}(n) \approx 2.22$ bits and the total NML codelength is approximately 4.13 bits.

For the original NML distribution with $v(\theta) = 1$, Rissanen [25] showed that

$$(45) \qquad -\log p_{\mathrm{NML}}(\mathbf{y}|\mathcal{M}) = -\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathcal{M}) + \log \int_{\Theta} \sqrt{|J_1(\boldsymbol{\theta})|} + \frac{p}{2} \log \left(\frac{n}{2\pi}\right) + o(1)$$

where $J_1(\cdot)$ is the per-sample Fisher information matrix. Additionally, Rissanen proved that the NML codelength reduces to the well-known Bayesian information criterion (BIC) in the limit as the sample size $n \to \infty$ and demonstrated that the asymptotic approximation is accurate for large sample sizes. Other, somewhat shaper, approximations to this NML codelength exist, see for example, Mera et al. [19]. Rissanen's asymptotic expansion was recently generalized to include arbitrary luckiness functions and a novel methodology for evaluation of the general, non-asymptotic NML codelength for exponential family models [34].

Although there are many similarities between MML and MDL, there exist some important differences in their inherent inference philosophies. Unlike MML, MDL is non-Bayesian and does not allow for the use of prior information in inference. While MML provides new means of parameter estimation, MDL was originally based on the maximum likelihood estimator and aimed to discover the best model class, rather than the best single model, for a given data set. However, the recent inclusion of luckiness functions in the NML code allows for new MDL-based penalized maximum likelihood estimation procedures. Lastly, the NML codelength is not a two-part code as Strict MML and is instead derived to minimize the worst-case excess codelength relative to an ideal code. For a more detailed discussion of MML and MDL, we recommend [3] and [35] (pp. 413–415).

References

1. E. Agrell and T. Eriksson, *Optimization of lattices for quantization*, IEEE Transactions on Information Theory **44** (1998), no. 5, 1814–1828.
2. V. Balasubramanian, *MDL, Bayesian inference, and the geometry of the space of probability distributions*, Advances in Minimum Description Length: Theory and Applications (I. J. Myung P. D. Grünwald and M. A. Pitt, eds.), MIT Press, 2005, pp. 81–99.
3. Rohan A. Baxter and J.J. Oliver, *MDL and MML: Similarities and differences*, Tech. Report TR 207, Department of Computer Science, Monash University, 1994.
4. M. J. Bayarri and G. Garcia-Donato, *Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing*, Journal of the Royal Statistical Society (Series B) **70** (2008), no. 5, 981–1003.
5. E. T. Bell, *Exponential numbers*, The American Mathematical Monthly **41** (1934), no. 7, 411–419.
6. Vittoria Bruni, Maria Lucia Cardinali, and Domenico Vitulano, *A short review on minimum description length: An application to dimension reduction in PCA*, Entropy **24** (2022), no. 2, 269.
7. J. H. Conway and N. J. A. Sloane, *Sphere packing, lattices and groups*, third ed., Springer-Verlag, December 1998.
8. Steven de Rooij and Peter Grünwald, *An empirical study of minimum description length model selection with infinite parametric complexity*, Journal of Mathematical Psychology **50** (2006), no. 2, 180–192.

9. James G. Dowty, *SMML estimators for 1-dimensional continuous data*, The Computer Journal **58** (2015), no. 1, 126–133.

10. Graham E. Farr and C. S. Wallace, *The complexity of strict minimum message length inference*, Computer Journal **45** (2002), no. 3, 285–292.

11. Mithat Gönen, Wesley O Johnson, Yonggang Lu, and Peter H Westfall, *The Bayesian two-sample t test*, The American Statistician **59** (2005), no. 3, 252–257.

12. Quentin F. Gronau, Alexander Ly, and Eric-Jan Wagenmakers, *Informed Bayesian t-tests*, The American Statistician **74** (2019), no. 2, 137–143.

13. Peter Grünwald and Teemu Roos, *Minimum description length revisited*, International Journal of Mathematics for Industry **11** (2019), no. 01, 1930001.

14. Robert E. Kass and Adrian E. Raftery, *Bayes factors*, Journal of the American Statistical Association **90** (1995), no. 430, 773–795.

15. Riko Kelter, *Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality*, Computational Statistics **36** (2020), no. 2, 1263–1288.

16. Petri Kontkanen and Petri Myllymäki, *A linear-time algorithm for computing the multinomial stochastic complexity*, Information Processing Letters **103** (2007), no. 6, 227–233.

17. Robin Lamarche-Perrin, Yves Demazeau, and Jean-Marc Vincent, *A generic algorithmic framework to solve special versions of the set partitioning problem*, 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, IEEE, 2014.

18. Enes Makalic and Daniel F. Schmidt, *Minimum message length analysis of the Behrens–Fisher problem*, Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence, Springer Berlin Heidelberg, 2013, pp. 250–260.

19. Bruno Mera, Paulo Mateus, and Alexandra M. Carvalho, *On the minmax regret for statistical manifolds: the role of curvature*, 2020.

20. Tomasz Michalak, Talal Rahwan, Edith Elkind, Michael Wooldridge, and Nicholas R. Jennings, *A hybrid exact algorithm for complete set partitioning*, Artificial Intelligence **230** (2016), 14–50.

21. Ingram Olkin and John W. Pratt, *Unbiased estimation of certain correlation coefficients*, The Annals of Mathematical Statistics **29** (1958), no. 1, 201–211.

22. Bo Peng and Min Wang, *Objective Bayesian testing for the correlation coefficient under divergence-based priors*, The American Statistician **75** (2021), no. 1, 41–51.

23. Talal Rahwan, Tomasz P. Michalak, Michael Wooldridge, and Nicholas R. Jennings, *Coalition structure generation: A survey*, Artificial Intelligence **229** (2015), 139–174.

24. Jorma Rissanen, *Modeling by shortest data description*, Automatica **14** (1978), no. 5, 465–471.

25. _____, *Fisher information and stochastic complexity*, IEEE Transactions on Information Theory **42** (1996), no. 1, 40–47.

26. _____, *Strong optimality of the normalized ML models as universal codes and information in data*, IEEE Transactions on Information Theory **47** (2001), no. 5, 1712–1717.

27. _____, *Information and complexity in statistical modeling*, first ed., Information Science and Statistics, Springer, 2007.

28. _____, *Optimal estimation*, Information Theory Newsletter **59** (2009), no. 3, 1–20.

29. Jorma Rissanen and Teemu Roos, *Conditional NML universal models*, Proc. 2007 Information Theory and Applications Workshop (ITA-07),, IEEE Press, 2007, (Invited Paper), pp. 337–341.

30. Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson, *Bayesian t tests for accepting and rejecting the null hypothesis*, Psychonomic Bulletin & Review **16** (2009), no. 2, 225–237.

31. Tuomas Sandholm, Kate Larson, Martin Andersson, Onn Shehory, and Fernando Tohmé, *Coalition structure generation with worst case guarantees*, Artificial Intelligence **111** (1999), no. 1-2, 209–238.

32. Daniel F. Schmidt and Enes Makalic, *MML invariant linear regression*, The 22nd Australasian Joint Conference on Artificial Intelligence (Melbourne, Australia), 2009, pp. 312–321.

33. Gideon Schwarz, *Estimating the dimension of a model*, The Annals of Statistics **6** (1978), no. 2, 461–464.

34. Atsushi Suzuki and Kenji Yamanishi, *Exact calculation of normalized maximum likelihood code length using Fourier analysis*, 2018 IEEE International Symposium on Information Theory (ISIT), IEEE, 2018.

35. Chris S. Wallace, *Statistical and inductive inference by minimum message length*, first ed., Information Science and Statistics, Springer, 2005.

36. Chris S. Wallace and David M. Boulton, *An information measure for classification*, Computer Journal **11** (1968), no. 2, 185–194.

37. _____, *An invariant Bayes method for point estimation*, Classification Society Bulletin **3** (1975), no. 3, 11–34.

38. Chris S. Wallace and David L. Dowe, *Minimum message length and Kolmogorov complexity*, Computer Journal **42** (1999), no. 4, 270–283.

39. Chris S. Wallace and David L. Dowe, *Refinements of MDL and MML coding*, Computer Journal **42** (1999), no. 4, 330–337.

40. _____, *MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions*, Statistics and Computing **10** (2000), no. 1, 73–83.

41. Chris S. Wallace and Peter R. Freeman, *Estimation and inference by compact coding*, Journal of the Royal Statistical Society (Series B) **49** (1987), no. 3, 240–252.

42. _____, *Single-factor analysis by minimum message length estimation*, Journal of the Royal Statistical Society (Series B) **54** (1992), no. 1, 195–209.

43. Chris S. Wallace and J. D. Patrick, *Coding decision trees*, Machine Learning **11** (1993), no. 1, 7–22.

44. Min Wang and Guangying Liu, *A simple two-sample Bayesian t-test for hypothesis testing*, The American Statistician **70** (2016), no. 2, 195–201.

45. Ruud Wetzels, Dora Matzke, Michael D. Lee, Jeffrey N. Rouder, Geoffrey J. Iverson, and Eric-Jan Wagenmakers, *Statistical evidence in experimental psychology*, Perspectives on Psychological Science **6** (2011), no. 3, 291–298.