MML and Bayesianism: Similarities and Differences

(Introduction to Minimum Encoding Inference — Part II)

Tech Report 206
Department of Computer Science,
Monash University,
Clayton, Vic. 3168, Australia
December 1994 — Amended August 15th 1995

JONATHAN J. OLIVER (jono@cs.monash.edu.au)
ROHAN A. BAXTER (rohan@cs.monash.edu.au)
Computer Science Department
Monash University
Clayton, Victoria, 3168, AUSTRALIA

Abstract: This paper continues the introduction to minimum encoding inference given by Oliver and Hand. This series of papers were written with the objective of providing an introduction to this area for statisticians. We examine the relationship between Bayesianism and Minimum Message Length (MML) inference. We argue that MML augments Bayesian methods by providing a sound Bayesian method for point estimation which is invariant under non-linear transformations. We explore the issues of invariance of estimators under non-linear transformations, the role of the Fisher Information matrix in MML inference, and the apparent similarity between MML and the adoption of a Jeffreys' Prior. We then compare MML to an approximate method of Bayesian Model Class Selection. Despite apparent similarities in their expressions, the properties of the two approaches can be different.

Contents

1 Introduction

This paper continues the introduction to minimum encoding inference given by Oliver and Hand [18]. This series of papers were written with the objective of providing an introduction for statisticians to this area. In this paper, we use the Wallace interpretation [26, 28, 29, 31, 33] of minimum encoding inference which he termed Minimum Message Length (MML). For details of Rissanen's approach (termed Minimum Description Length or MDL) see [23, 24, 25]. A comparison of MML and MDL can be found in Baxter and Oliver [3].

1.1 MML and Bayesian Data Analysis

Given a data set, D, there are a number of ways in which we may wish to use D. In this paper, we distinguish between the following uses of data analysis:

- Prediction — we may make predictions about future data using the assumption that our predictions will be improved by knowledge of D.
- Decision Making — we may attempt to make decisions based on D to maximise some utility function.
- Inference — we may infer things about the real world. This is often cast as parameter estimation and model selection.

1.2 A Bayesian Framework

In a Bayesian setting, we consider a parameterisation, θ, and examine the posterior probability distribution, $Prob(\theta \mid D)$. Bayes' theorem tells us:

$$\boxed{Prob(\theta \mid D) \;=\; \frac{Prob(\theta \;\&\; D)}{Prob(D)} \;=\; \frac{Prob(\theta) \;\times\; Prob(D \mid \theta)}{Prob(D)}} \qquad (1)$$

We write the prior probability distribution, $Prob(\theta)$ as $h(\theta)$, and the likelihood, $Prob(D \mid \theta)$, as $f(D \mid \theta)$. Since $Prob(D)$ is a constant, we view the posterior, $Prob(\theta \mid D)$, as being proportional to the product of the prior and likelihood:

$$Posterior(\theta) \;\propto\; h(\theta) \;\times\; f(D \mid \theta)$$

If we have a set of measurements $z_i \in D$, and we assume that each z_i is independent:

$$f(D \mid \theta) \;=\; \prod_{z_i \,\in\, D} f(z_i \mid \theta)$$

and therefore:

$$Posterior(\theta) \;\propto\; h(\theta) \;\times\; \prod_{z_i \,\in\, D} f(z_i \mid \theta)$$

In this paper, we focus on those problems where θ is continuous; in these problems the posterior density attaches zero probability to any given value of θ.

There is general agreement on how a posterior density may be used to achieve optimal predictions, and make optimal decisions [4, 20]. However, there is not general agreement on how (if at all) a posterior should be used to perform inference. Section 2 discusses some methods for using the posterior to give parameter estimates, and discuss problematic features of these estimation schemes. Section 3 discusses the failure of these estimators to be invariant under non-linear parameter transformations.

1.3 An MML Framework

In an MML setting, we calculate message lengths for the description of θ, and the description of the data (assuming the given value of θ):

$$\boxed{MessLen(D \;\&\; \theta) \;=\; MessLen(\theta) \;+\; MessLen(D \mid \theta)} \qquad (2)$$

We take the value of θ which minimises the total message length as the MML estimate.

To calculate the message length of the data given θ, we note that data values we are given come to some *accuracy of measurement* (AOM)[1]. For example, the height of a person may have been measured to within a 1 cm range. For a set of measurements $z_i \in D$, with each z_i independent, with a given AOM, we take

$$Prob(D \mid \theta) \approx \prod_{z_i \in D} f(z_i \mid \theta) AOM$$

If the AOM is constant over the range of data values, we may re-scale our measurements so that $AOM = 1$. Therefore[2]

$$\boxed{MessLen(D \mid \theta) = \sum_{z_i \in D} -\log\left(f(z_i \mid \theta)\right) \; nits} \qquad (3)$$

We calculate the message length of the parameter, by dividing a prior density $h(\theta)$ into cells, and associating a code word with each cell. The width of these cells is termed the *Accuracy of Parameter Value* (AOPV)[3]. We allow $AOPV_\theta$ to be a function of θ. $AOPV_\theta$ is used to approximate $Prob(\theta)$:

$$Prob(\theta) \approx h(\theta) \times AOPV_\theta$$

and hence

$$\boxed{MessLen(\theta) = -\log\left(h(\theta) \times AOPV_\theta\right)} \qquad (4)$$

In effect, the MML approach discretises the prior distribution into regions of size $AOPV_\theta$, as shown in Figure 1.

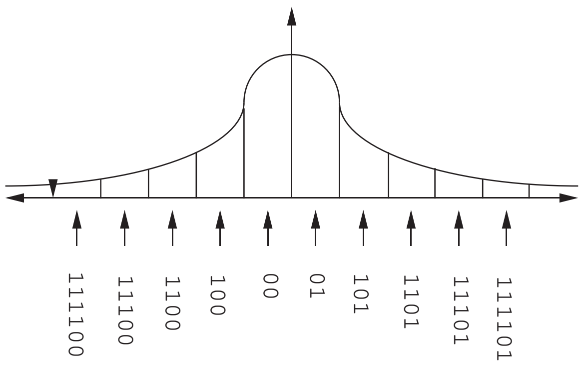

Figure 1: A Discretised Prior Density

Using MML for estimation requires us to determine $AOPV$ values in some sensible manner. We discuss how to determine optimal AOPVs in Section 4. Throughout this paper, we interpret MML in a Bayesian light. We find that MML offers Bayesianism a sound method for inference.

[1] AOMs are discussed in Section 2.7 of Oliver and Hand [18].
[2] For convenience, we calculate all message lengths in nits. We therefore use natural logarithms throughout this paper.
[3] The division of a prior density into cells and AOPVs are discussed in Section 3.1 of Oliver and Hand [18].

2 Bayesian Inference

Fundamentalist Bayesians reject attempts to summarise a posterior density by an estimate as being unsound (e.g., Neal [17, Chapter 1]). For them, the posterior density is a sufficient and satisfactory end result of inference. In some cases, we may be able to present the full posterior graphically. For multivariate densities, this is often infeasible in practice.

However, we often wish to summarise the posterior by using an estimate. In a traditional Bayesian framework, a single estimate may be found if there is a clearly defined and known loss function, $l(\theta,\ \hat{\theta})$. This function specifies the "loss" occurring if the estimate is used when the true value is not the estimate. The optimal value for an estimate is found by minimising the expected loss:

$$\int_{all\ \theta} l(\theta,\ \hat{\theta})\ Prob(\theta \mid D)\ d\theta$$

Three types of loss functions are commonly identified:

- A $0-1$ loss function: $l(\theta,\ \hat{\theta}) \propto 1 - Prob(B_\epsilon(\hat{\theta}))$
  where $B_\epsilon(\hat{\theta})$ is a ball of radius ϵ centered at $\hat{\theta}$ [4, Page 257].
- A quadratic loss function: $l(\theta,\ \hat{\theta}) \propto (\ \theta - \hat{\theta}\ )^2$
- A linear loss function: $l(\theta,\ \hat{\theta}) \propto \mid \theta - \hat{\theta} \mid$

These three loss functions lead to the following estimates [14]:

- $0-1$ loss — **Mode.** — The value which maximises the posterior density:
  $$\hat{\theta} = \max(Posterior(\theta))$$
- Quadratic loss — **Mean.**
  $$\hat{\theta} = \int Posterior(\theta)\ \theta\ d\theta$$
- Linear loss — **Median.** — The value of θ which satisfies:
  $$\int_{-\infty}^{\hat{\theta}} Posterior(\theta)\ d\theta = \int_{\hat{\theta}}^{\infty} Posterior(\theta)\ d\theta$$

However, estimates are sometimes required in circumstances when no loss function is available. For example, a physicist would like to know the estimated mass of an electron. The physicist will not necessarily know what uses the estimated mass will be put too.

2.1 Problems with using the Posterior to Perform Estimation

In this section, we shall identify problems with the three methods of estimation given in the previous section.

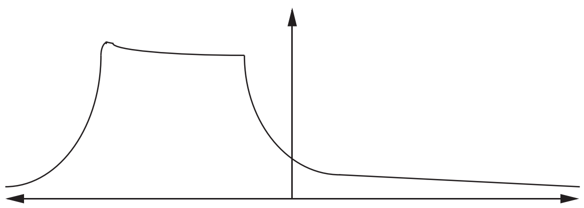

Figure 2: A posterior density with a plateau with a peak at one end

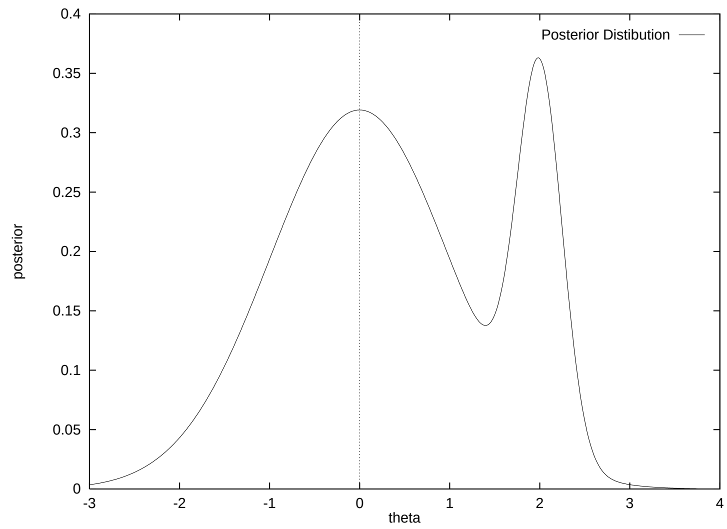

Figure 3: A posterior density with two peaks

- **Mode.** — Firstly, selecting the peak of the posterior is not invariant under non-linear transformations (discussed at length in Section 3). Secondly, we may find that regions where the posterior density is slightly greater than its surrounding region may be selected (shown in Figure 2). We typically would not want to select $\hat{\theta}$ as being the top of the plateau in Figure 2, since there is more probability associated with a region around an estimate near the center of the plateau. Thirdly, a mode of the posterior which contains a considerable probability mass may be lower than a mode containing less probability mass. For example, the mode centered at $\theta = 0$ in Figure 3 contains 80% of the probability mass, while the (higher) mode centered at $\theta = 2$ contains 20% of the probability mass.

- **Mean.** — Firstly, selecting the mean of the posterior is not invariant under non-linear transformations. Secondly, we may find that the mean of the posterior is in a region with little probability associated with it, as shown in Figure 4[4]. This second criticism may appear artificial. However, there are models where symmetric multimodal posteriors exist. For example, in Factor Analysis [32], a sign ambiguity means that posteriors will always be symmetric about 0. Bayesian Factor Analysis has therefore tended to use the mode as an estimate [1, 21].

- **Median.** — Selecting the median of the posterior has the property that it IS invariant under non-linear transformations. However, it does not generalise to two or more parameters, and may select a region of the posterior with little probability associated with it (Figure 4).

[4] If the posterior is multimodal we may question the attempt to summarise it with a single estimator.

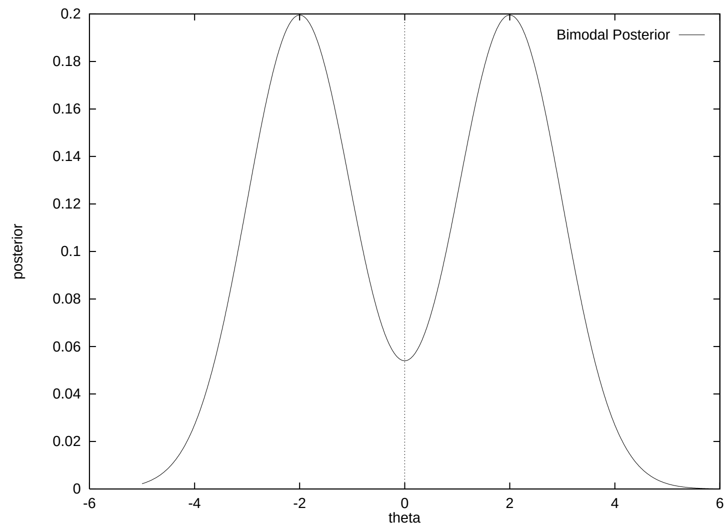

Figure 4: A symmetric posterior density with two peaks

3 Invariance under Transformations

When performing estimation, it is important to consider the effect of transformations made to the way we measure the data or parameters. For example if we measure the temperature for a month and calculate the average temperature, then we demand that the average be the same whether we measure the temperature in celsius or fahrenheit. This is an example of a linear transformation:

$$Temperature_C(day) \;=\; \frac{5}{9}(Temperature_F(day) - 32)$$

Consider the effect of transforming a parameter space such that

$$\theta' \;=\; Transform(\theta)$$

Let $\hat{\theta}$ be the estimate in the original parameter space, and $\hat{\theta'}$ be the estimate in the transformed parameter space. An *invariant* estimator satisfies:

$$\hat{\theta'} \;=\; Transform(\hat{\theta})$$

For example, we may perform an experiment to determine information about a particle (of known mass m) after a specified interaction. We may then analyse the results to estimate the speed, $\hat{v}$ of the particle, and the kinetic energy, $\hat{E_k}$ of the particle. Given that we used the same data in both scenarios, our estimator should conclude that:

$$\hat{E_k} \;=\; \frac{1}{2}\; m\; \hat{v}^2$$

Many estimators are invariant under linear transformations (such as the temperature example), but are not invariant under non-linear transformations (such as the particle example).

3.1 An Example Non-linear Transformation

In this section, we consider the effect of a non-linear transformation applied to a posterior density. Consider the problem of estimating a magnetic field strength from a set of compass directions. There are two parameterisations often used for a two dimensional vector (a magnetic field strength in this case):

- (κ, μ) — in polar co-ordinates. The length of the vector is κ, and its direction is μ.
- $x\tilde{\imath} \;+\; y\tilde{\jmath}$ — in x-y co-ordinates.

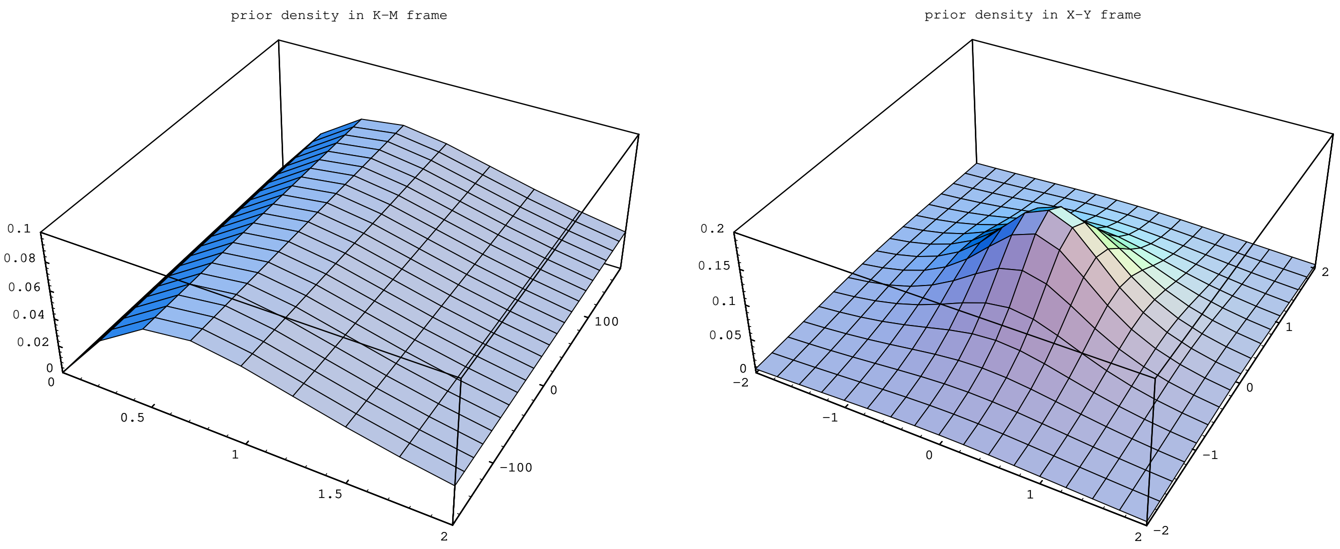

Figure 5: The Prior Density in Polar and x-y Co-ordinates

We may map one parameterisation onto the other by using the transformation: $x = \kappa\, cos(\mu)$, $y = \kappa\, sin(\mu)$. We now give a derivation of the posterior distribution using a von Mises likelihood function for the a set of 10 independent compass directions:

$$D = \{\omega_1,\ \omega_2,\ \ldots,\ \omega_{10}\} = \{279^o,\ 143^o,\ 307^o,\ 153^o,\ 35^o,\ 203^o,\ 325^o,\ 45^o,\ 20^o,\ 74^o\}$$

We derive expressions for the posterior in both polar co-ordinates, and in x-y co-ordinates. We then use the derived expressions to estimate the pair (κ, μ), and the pair (x, y).

The probability density function of a von Mises distribution on a circle is:

$$f_{KM}(\omega) = \frac{1}{2\pi I_0(\kappa)}\, e^{\kappa\ cos(\omega-\mu)} \tag{5}$$

$$f_{XY}(\omega) = \frac{1}{2\pi I_0(\sqrt{x^2\ +\ y^2})}\, e^{\sqrt{x^2\ +\ y^2}\ cos(\omega - tan^{-1}(\frac{y}{x}))} \tag{6}$$

where $I_0(\kappa)$ is the modified Bessel function.

We wish to adopt a specific prior, and investigate the effect of reparameterisations on the posterior. Wallace and Dowe [30, Page 7] found $h_3 = \frac{\kappa}{(1\ +\ \kappa^2)^{\frac{3}{2}}}$ was an appropriate prior for a range of problems. If we assume a uniform prior distribution over μ, then we get:

$$h_{KM}(\kappa,\ \mu) = \frac{\kappa}{2\pi(1\ +\ \kappa^2)^{\frac{3}{2}}} \tag{7}$$

$$h_{XY}(x,\ y) = \frac{1}{2\pi(1\ +\ x^2\ +\ y^2)^{\frac{3}{2}}} \tag{8}$$

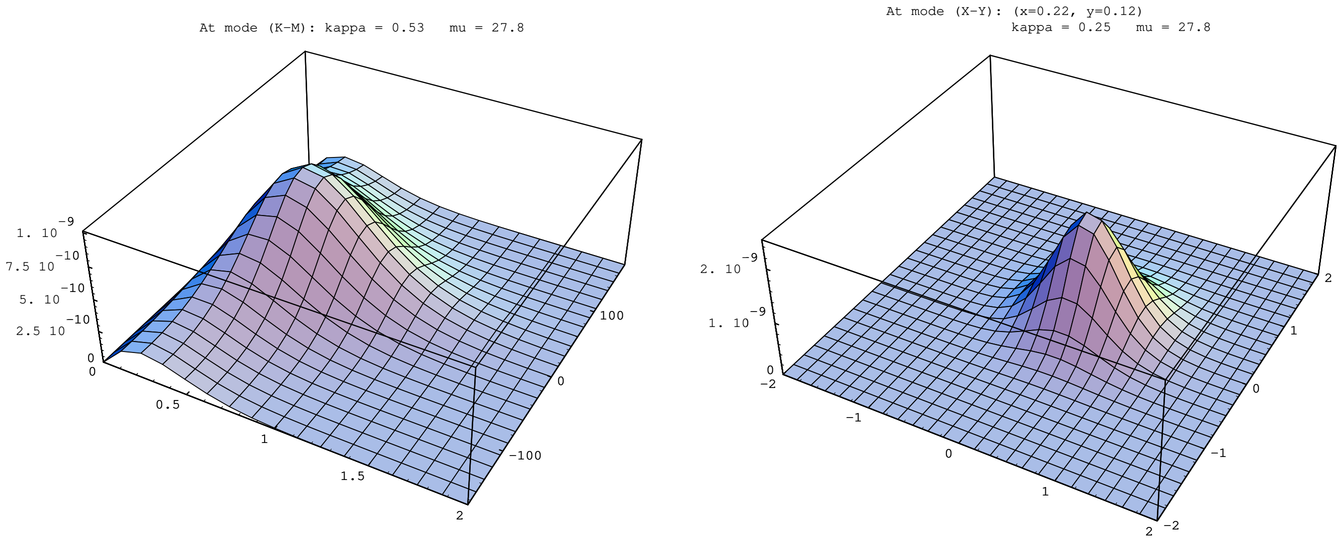

Figure 6: The Posterior Density in Polar and x-y Co-ordinates

These prior densities are depicted in Figure 5. We note that to transform from the (κ, μ) frame to the (x, y) frame, we had to divide $h_{KM}(\kappa,\ \mu)$ by the Jacobian of the transformation[5] $(J\ =\ \kappa)$[6]:

$$h_{XY}(x,\ y)\ =\ \frac{h_{KM}(\kappa,\ \mu)}{\kappa}$$

The posterior distributions are:

$$Posterior_{KM}(\kappa,\ \mu)\ \propto\ h_{KM}(\kappa,\ \mu)\ \times\ \prod_{i=1}^{10} f_{KM}(\omega_i) \tag{9}$$

$$Posterior_{XY}(x,\ y)\ \propto\ h_{XY}(x,\ y)\ \times\ \prod_{i=1}^{10} f_{XY}(\omega_i) \tag{10}$$

These posterior densities are depicted in Figure 6. We find that the mode of the posterior has moved from $(\kappa = 0.53,\ \mu = 27.8)$ to $(x = 0.22,\ y = 0.12)$ which is equivalent to $(\kappa = 0.25,\ \mu = 27.8)$.

[5] When we transform a prior distribution from one frame to another, we are required to divide by the Jacobian of the transformation, to keep the integral of the prior with a given region constant. For example, consider a prior for temperatures which is uniform in the range $[0^oC,\ 100^oC]$, so $h_C(temp)\ =\ \frac{1}{100}$. If we transform this prior to Fahrenheit temperatures, then we have a uniform prior in the range $[32^oC,\ 212^oC]$ with value $h_F(temp)\ =\ \frac{1}{180}$. The Jacobian of this transformation is $J\ =\ \frac{9}{5}$, and we find $h_C(temp)\ =\ \frac{5}{9}\ \times\ h_F(temp)$.

[6] The Jacobian is calculated in the following way:

$$J\ =\ det\begin{bmatrix} \frac{\partial x}{\partial \kappa} & \frac{\partial x}{\partial \mu} \\ \frac{\partial y}{\partial \kappa} & \frac{\partial y}{\partial \mu} \end{bmatrix}\ =\ det\begin{bmatrix} cos\mu & -\kappa\ sin\mu \\ sin\mu & \kappa\ cos\mu \end{bmatrix}\ =\ \kappa(cos^2\mu\ +\ sin^2\mu)\ =\ \kappa$$

4 MML Estimators

In this section, we discuss MML estimators. We find that the MML framework for determining estimates establishes a method for determining AOPVs.

4.1 The Original MML Method

In [18] we used the Original MML Method (offered by Wallace and Boulton [28]) for determining the AOPVs of μ and σ for a normal distribution. We did this by:

1. Constructing an expression for the message length of a message taking the form:

$$< \mu > < \sigma > < z_1,\ z_2,\ \ldots\ z_N >$$

   in terms of $AOPV_\mu$, $AOPV_\sigma$, N and statistics of the z_i's:[7]

$$MessLen(D \,\&\, \theta) \;=\; \log \frac{range_\mu}{AOPV_\mu} \;+\; \log \frac{range_\sigma}{AOPV_\sigma} \;+\; N\, \log\, \frac{\bar{s}\,\sqrt{2\pi}}{AOM} \;+\; N\, \frac{s^2 \;+\; \frac{\bar{s}^2}{N}}{2\,\bar{s}^2} \qquad (11)$$

   where s^2 is the biased sample variance, where $\bar{s}^2$ is the unbiased sample variance, and $range_\mu$ and $range_\sigma$ are the ranges of possible μ and σ values, respectively.

2. Differentiating Equation (11) w.r.t. $AOPV_\mu$ to find the optimal $AOPV_\mu$.
3. Differentiating Equation (11) w.r.t. $AOPV_\sigma$ to find the optimal $AOPV_\sigma$.

The optimal AOPVs were:

$$AOPV_\mu \;=\; \sigma\sqrt{\frac{12}{N}} \qquad AOPV_\sigma \;=\; \sigma\sqrt{\frac{6}{N-1}}$$

While such an approach seemed reasonable for this problem, it has the following problems:

- The use of successive differentiation does not guarantee the optimal choice for the pair $(AOPV_\mu,\ AOPV_\sigma)$.
- The method does not generalise gracefully to many dimensions. The derivation for the "optimal" values for two variables using the successive differentiation approach [28] was an involved calculation.

4.2 A Bayesian Interpretation of the Original MML Method

The Original MML Method can be interpreted as a Bayesian method for point estimation. In the example discussed in Section 4.1 (models consisting of normal curves), the model space can be represented as a two dimensional space with $\mu \in (-\infty,\ \infty)$, and $\sigma \in (0,\ \infty)$. The Original MML Method partitions this space into rectangles as shown in Figure 7. The height of each rectangle is $AOPV_\sigma$, and the width is $AOPV_\mu$.

We may rewrite Equation (11) as:

$$MessLen(D \,\&\, \theta) \;=\; -\,\log(h(\mu,\ \sigma)\ \times\ AOPV_\mu\ \times\ AOPV_\sigma) - \log f(D \mid \mu,\ \sigma) + Imprecision\ Term \quad (12)$$

where $h(\mu,\ \sigma) \;=\; \frac{1}{range_\mu\ \times\ range_\sigma}$, and the $Imprecision\ Term$ is added since we cannot state the data D in $-\ \log f(D \mid \mu,\ \sigma)$ nits, since we stated the parameter values imprecisely.

We take as the point estimate for the parameter, $\theta \;=\; (\mu,\ \sigma)$, the midpoint of the rectangle with the maximum posterior probability.

4.3 Generalising the Original MML Method

We can generalise and improve the Original MML Method (Section 4.1) by using the approach taken by Wallace and Freeman [31]. Rather than successively optimising a set of parameters in turn with a particular distribution in mind, we consider a more general problem. We assume the distribution $f(.)$ to be a regular distribution, and attempt to optimise the parameters together.

In Section 4.4 we consider this more general problem for a single parameter. Section 5 extends this to a vector of parameters.

[7] Equation (11) is Equation (14) from Oliver and Hand [18] rewritten in nits.

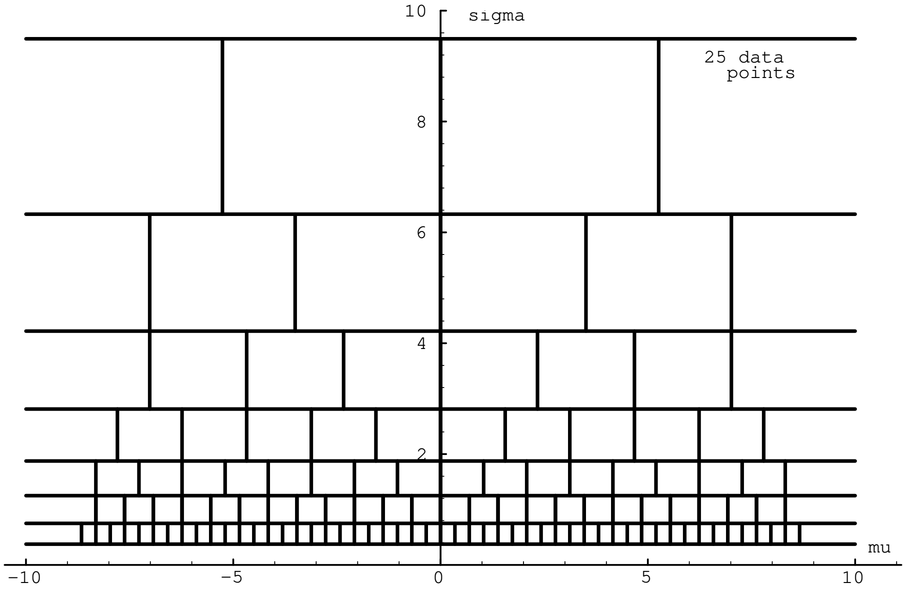

Figure 7: Coding Parameters with Rectangles

4.4 Estimating a Single Parameter Using MML

We estimate a single parameter, θ from some data D with prior density $h(\theta)$, likelihood $f(D \mid \theta)$ and negative log-likelihood, $L(\theta) = -\log f(D \mid \theta)$. We assume that θ is described to accuracy $AOPV_\theta = A(\theta)$.

4.4.1 Deriving an Expression for the Message Length

We put $A(\theta)$ in Equation (4) (from Section 1.3).

$$MessLen(\theta) = -\log(A(\theta) \times h(\theta)) = -\log A(\theta) - \log h(\theta) \tag{13}$$

We define $MessLen_{Expected}(D \mid \theta)$ to be the expected message length[8] in the region $(\theta - \frac{A(\theta)}{2},\ \theta + \frac{A(\theta)}{2})$:

$$MessLen_{Expected}(D \mid \theta) = \frac{1}{A(\theta)} \int_{\frac{-A(\theta)}{2}}^{\frac{A(\theta)}{2}} MessLen(D \mid \theta + x)\, dx$$

x is the discrepancy between θ (a point estimate) and the value of the parameter used to code the data. We assume that x is uniform in the region $(-\frac{A(\theta)}{2},\ \frac{A(\theta)}{2})$.
We write $MessLen_{Expected}(D \mid \theta)$ in terms of the negative log-likelihood, using Equation (3) (from Section 1.3):

$$MessLen_{Expected}(D \mid \theta) = \frac{1}{A(\theta)} \int_{\frac{-A(\theta)}{2}}^{\frac{A(\theta)}{2}} L(\theta + x)\, dx \tag{14}$$

Substituting Equation (13) and Equation (14) into Equation (2) (from Section 1.3) gives:

$$MessLen(\theta \;\&\; D) = -\log A(\theta) - \log h(\theta) + \frac{1}{A(\theta)} \int_{\frac{-A(\theta)}{2}}^{\frac{A(\theta)}{2}} L(\theta + x)\, dx$$

We expand out the integral in terms of the Taylor's expansion (to terms of the second power):

$$MessLen(\theta \;\&\; D) \approx -\log A(\theta) - \log h(\theta) + \frac{1}{A(\theta)} \int_{\frac{-A(\theta)}{2}}^{\frac{A(\theta)}{2}} \left(L(\theta) + x\frac{\partial L(\theta)}{\partial \theta} + \frac{x^2}{2}\frac{\partial^2 L(\theta)}{\partial \theta^2} \right) dx$$

[8] An argument for this definition is given in Section 3.3.1. of Oliver and Hand [18].

We note that:

$$\frac{1}{A(\theta)} \int_{\frac{-A(\theta)}{2}}^{\frac{A(\theta)}{2}} L(\theta)\ dx \ = \ L(\theta)$$

and

$$\frac{1}{A(\theta)} \int_{\frac{-A(\theta)}{2}}^{\frac{A(\theta)}{2}} x \frac{\partial L(\theta)}{\partial \theta}\ dx \ = \ 0$$

Therefore

$$\begin{aligned} MessLen(\theta\ \&\ D) &\approx -\log\ A(\theta) \ - \ \log\ h(\theta) \ + \ L(\theta) \ + \ \frac{\frac{\partial^2 L(\theta)}{\partial \theta^2}}{2\ A(\theta)} \int_{\frac{-A(\theta)}{2}}^{\frac{A(\theta)}{2}} x^2\ dx \\ &\approx -\log\ A(\theta) \ - \ \log\ h(\theta) \ + \ L(\theta) \ + \ \frac{\frac{\partial^2 L(\theta)}{\partial \theta^2}}{2\ A(\theta)} \frac{A(\theta)^3}{12} \\ &\approx -\log\ A(\theta) \ - \ \log\ h(\theta) \ + \ L(\theta) \ + \ \frac{A(\theta)^2 \frac{\partial^2 L(\theta)}{\partial \theta^2}}{24} \end{aligned}$$

4.4.2 Finding the Optimal AOPV

We select that value for $A(\theta)$ which minimises the message length by setting $\frac{\partial MessLen(\theta\ \&\ D)}{\partial A(\theta)} \ = \ 0.$

$$\frac{\partial MessLen(\theta\ \&\ D)}{\partial A(\theta)} \ = \ \frac{-1}{A(\theta)} \ + \ \frac{A(\theta) \frac{\partial^2 L(\theta)}{\partial \theta^2}}{12} \ = \ 0 \qquad (15)$$

The optimal value for $A(\theta)$ is therefore:

$$A(\theta) \ = \ \sqrt{\frac{12}{\frac{\partial^2 L(\theta)}{\partial \theta^2}}} \qquad (16)$$

We wish to use the optimal value of the AOPV to construct a *code dictionary*. A code dictionary[9] gives a code word for each model (in this case a parameter value) under consideration. There is a problem with creating a code dictionary using the optimal AOPV value given by Equation (16). The optimal AOPV value is dependent on the data (since $L(\theta) \ = \ -\log\ f(D \mid \theta)$ is a function of the data), and hence a message of this form will not be decodeable.

4.4.3 Creating a Decodeable Message

One apparent solution to this problem is to have a 3-part code which takes the form:

$< AOPV_\theta > < \theta > < D >$

This could use a prior over $AOPV_\theta$ values. It seems difficult to specify and justify such a prior. In addition we must also state the precision to which we state $AOPV_\theta$. Thus we have a message of the form:

$< AOPV_{AOPV_\theta} > < AOPV_\theta > < \theta > < D >$

This apparently leads to an infinite regress since we now require a prior over $AOPV_{AOPV_\theta}$ values, and we must also state the precision to which we state $AOPV_{AOPV_\theta}$.

An alternative suggested by Wallace and Freeman [31] is to use the "average" optimal AOPV. Given θ, we find each data set D has a different optimal AOPV. Thus, each θ has a set of optimal AOPV values. We could define the "average" AOPV as being:

$$Average(AOPV_\theta) \ = \ \int f(D \mid \theta)\ A(\theta)\ dz$$

The sub-expression in $A(\theta)$ which is a function of the data is the Observed Fisher Information, $M(\theta) \ = \ \frac{\partial^2 L(\theta)}{\partial \theta^2}$. We may create a decodeable message by using $AOPV_\theta \ = \ A'(\theta)$ where $A'(\theta)$ is the value given by Equation (16) with the Observed Fisher Information, $M(\theta)$, replaced by the Expected Fisher Information, $F(\theta)$,

[9] Code dictionaries are discussed in Section 2 of Oliver and Hand [18].

We therefore replace

$$M(\theta) \;=\; \frac{\partial^2 L(\theta)}{\partial \theta^2} \qquad \Rightarrow \qquad F(\theta) \;=\; \int f(D \mid \theta) \frac{\partial^2 L(\theta)}{\partial \theta^2} \; dz$$

Thus, we can construct a code using $AOPV_\theta$:

$$A'(\theta) \;=\; \sqrt{\frac{12}{F(\theta)}} \tag{17}$$

Using this approximation allows us to construct a coding scheme with decodeable messages. $F(\theta)$ is not dependent upon the data, and both the receiver and sender of a message can determine $F(\theta)$ (and hence construct a code dictionary) before the message is transmitted.

Taking this expectation is an essential difference between the MML method and some Bayesian methods (discussed in Section 6). This issue is discussed in detail in Section 3.3 of Wallace and Boulton [29].

4.4.4 Approximating the Optimal Message Length

Substituting the optimal $AOPV_\theta$ from Equation (17) into Equation (15) gives us:

$$MessLen(\theta \;\&\; D) \;\approx\; -\log \sqrt{\frac{12}{F(\theta)}} \;-\; \log\; h(\theta) \;+\; L(\theta) \;+\; \frac{\frac{12}{F(\theta)} F(\theta)}{24}$$

Simplifying,

$$\boxed{MessLen(\theta \;\&\; D) \;\approx\; \frac{1}{2} \log\; \frac{1}{12} \;+\; \frac{1}{2} \log\; F(\theta) \;-\; \log\; h(\theta) \;+\; L(\theta) \;+\; \frac{1}{2}} \tag{18}$$

5 The Wallace-Freeman MML Method

In this section, we generalise the approach from Section 4.4 to the d dimensional case. This is an expanded derivation of the Wallace-Freeman MML Method presented to the Royal Statistical Society [31] (also given in Wallace and Dowe [30]).

We consider estimating a vector of parameters, θ from some data D with prior density $h(\theta)$, likelihood $f(D \mid \theta)$ and negative log-likelihood, $L(\theta) \;=\; -\log\; f(D \mid \theta)$. We assume that the d dimensional parameter space is partitioned into regions of volume $V \;=\; AOPV_\theta$.

5.0.5 Deriving an Expression for the Message Length

We define the d dimensional equivalent of Equation (14):

$$MessLen(\theta \mid D) \;=\; \frac{1}{V} \int_V MessLen(\theta + \bar{x}) \; dv$$

Therefore

$$\begin{aligned} MessLen(\theta \;\&\; D) \;&\approx\; -\log V h(\theta) \;+\; \frac{1}{V} \int_V \left(\; L(\theta) \;+\; \bar{x} \frac{\partial L(\theta)}{\partial \theta} \;+\; \frac{1}{2} \bar{x}^T \frac{\partial^2 L(\theta)}{\partial \theta^2} \bar{x} \; \right) \; dv \\ &\approx\; -\log V h(\theta) \;+\; \frac{1}{V} \left(\; \int_V L(\theta) dv \;+\; \int_V \bar{x} \frac{\partial L(\theta)}{\partial \theta} dv \;+\; \frac{1}{2} \; \int_V \bar{x}^T \frac{\partial^2 L(\theta)}{\partial \theta^2} \bar{x} dv \right) \end{aligned}$$

$$MessLen(\theta \;\&\; D) \;\approx\; -\log V h(\theta) \;+\; L(\theta) \;+\; \frac{1}{2} \, \frac{1}{V} \int_V \; \bar{x}^T \frac{\partial^2 L(\theta)}{\partial \theta^2} \bar{x} \; dv \tag{19}$$

We wish to find an expression for V which minimises Equation (19). To minimise it, we begin by examining how we may minimise the integral:

$$I \;=\; \frac{1}{2V} \int_V \left(\; \bar{x}^T \frac{\partial^2 L(\theta)}{\partial \theta^2} \bar{x} \; \right) \; dv$$

To make the message decodeable, we make the approximation of replacing $\frac{\partial^2 L(\theta)}{\partial \theta^2}$ with the Fisher information:

$$I \;\approx\; \frac{1}{2V} \int_V \;(\; \bar{x}^T \; F(\theta) \; \bar{x} \;) \;\; dv \tag{20}$$

To evaluate this integral, we apply a reparameterization transformation B^{-1} such that:

$$\bar{y} \;=\; B^{-1} \bar{x}$$

In Appendix 1, we construct B^{-1} such that:

$$\bar{x}^T \; F(\theta) \; \bar{x} \;=\; \bar{y}^T \; \bar{y}$$

Let $g(\phi)$ be the transformed prior density, and let volume V (in the x-space) map onto volume U (in the ϕ-space). In ϕ-space, the message length is:

$$\begin{aligned} MessLen(\theta \;\&\; D) \;\; &\approx \;\; -\log(U \; g(\phi)) \;+\; L(\theta) \;+\; \frac{1}{2} \, \frac{1}{U} \int_U \;(\; \bar{y}^T \; \bar{y} \;) \;\; du \\ &\approx \;\; -\log(U \; g(\phi)) \;+\; L(\theta) \;+\; \frac{1}{2} \; E(\bar{y}^T \; \bar{y}) \end{aligned}$$

The expected value of $\bar{y}^T \; \bar{y}$ is expressed in terms of the d dimensional optimal quantizing lattice constant, κ_d:

$$E(\bar{y}^T \; \bar{y}) \;=\; d \; \kappa_d \; U^{\frac{2}{d}} \tag{21}$$

where $\kappa_1 \;=\; \frac{1}{12}$ (since $\int_{\frac{-1}{2}}^{\frac{1}{2}} y^2 \; dy \;=\; \frac{1}{12}$) and $\kappa_2 \;=\; \frac{5}{36\,\sqrt{3}}$. Conway and Sloane [6] give bounds on the d dimensional optimal quantizing lattice constants, κ_d. Therefore

$$MessLen(\theta \;\&\; D) \;\approx\; -\log(U \; g(\phi)) \;+\; L(\theta) \;+\; \frac{d}{2} \; \kappa_d \; U^{\frac{2}{d}} \tag{22}$$

5.0.6 Finding the Optimal AOPV

To find the optimal AOPV, V, we differentiate Equation (22) w.r.t. U:

$$\frac{\partial MessLen(\theta \;\&\; D)}{\partial U} \;=\; -\frac{1}{U} \;+\; \frac{d}{2} \; \kappa_d \; \frac{2}{d} \; U^{\frac{2}{d}-1} \;=\; -\frac{1}{U} \;+\; \kappa_d \; U^{\frac{2}{d}-1}$$

Setting $\frac{\partial MessLen(\theta \;\&\; D)}{\partial U}$ to 0 gives:

$$\begin{aligned} \frac{1}{U} \;\; &= \;\; \kappa_d \; U^{\frac{2}{d}-1} \\ U \;\; &= \;\; \kappa_d^{-\frac{d}{2}} \end{aligned}$$

5.0.7 Approximating the Optimal Message Length

Substituting the optimal value for U into Equation (21) gives

$$E(\bar{y}^T \; \bar{y}) \;=\; d$$

Therefore the message length is:

$$MessLen(\theta \;\&\; D) \;\approx\; -\log(U \; g(\phi)) \;+\; L(\theta) \;+\; \frac{d}{2}$$

We wish to express $g(\phi)$ in terms of $h(\theta)$

$$g(\phi) \;=\; h(\theta) \; \frac{dV}{dU}$$

To do this, we note that

$$U = Jacobian(B^{-1})\ V$$

and therefore:

$$\frac{dV}{dU} = \frac{1}{Jacobian(B^{-1})}$$

We establish in Appendix 2 that

$$Jacobian(B^{-1}) = \sqrt{det(F(\theta))}$$

Therefore

$$MessLen(\theta\ \&\ D) \approx -\log(U) - \log(\frac{h(\theta)}{\sqrt{det(F(\theta))}}) + L(\theta) + \frac{d}{2}$$

Substituting the optimal value for U gives:

$$\boxed{MessLen(\theta\ \&\ D) \approx \frac{d}{2}\log(\kappa_d) - \log(h(\theta)) + \frac{1}{2}\log(det(F(\theta))) + L(\theta) + \frac{d}{2}} \tag{23}$$

We define the MML estimate, $\hat{\theta}_{MML}$, as the parameter value which minimises Equation (23).

5.1 Coding Data Using the Wallace-Freeman MML Method

The Wallace-Freeman MML Method constructs a code dictionary using the following process:

1. We partition up the parameter space. This can be done by examining the likelihood function, and determining the optimal volume.

   $$AOPV_\theta = V = \frac{\kappa_d^{-\frac{d}{2}}}{\sqrt{det(F(\theta))}}$$

   Note that the partitioning is independent of the prior distribution.

2. We assign each region, R, a prior probability, P_R, which is the integrated prior for that region. If the prior, $h(\theta)$ is suitably flat, then we may set

   $$P_{\hat{\theta}} = \int_R h(\theta)\ d\theta = AOPV_\theta\ h(\hat{\theta})$$

   where $\hat{\theta}$ is some point in R.

3. We assign a code word for that region which is of length $-\log(P_{\hat{\theta}})$.

Given a parameter estimate $\hat{\theta}$, we send a message of length $-\log(P_{\hat{\theta}})$ to describe $\hat{\theta}$. We may then send the data in (approximately) $L = -\log f(D \mid \theta)$ nits.

5.2 A Bayesian Interpretation of the Wallace-Freeman MML Method

We may rewrite Equation (23) as:

$$MessLen(D\ \&\ \theta) = -\log(h(\theta)AOPV_\theta) - \log f(D \mid \theta) + Imprecision\ Term \tag{24}$$

where the *Imprecision Term* is added because we stated the parameter values imprecisely. We are therefore selecting the point estimate with an associated region of maximal posterior probability.

5.3 Differences between MML and the Adoption of a Jeffreys' Prior

It may appear that the Wallace-Freeman MML approach is related to maximising the posterior when we use a Jeffreys' Prior [11, 12] of the form:

$$h(\theta) \propto \sqrt{det(F(\theta))}$$

The Wallace-Freeman approach is different from the use of a Jeffreys' Prior in the following ways:

- The Wallace-Freeman approach allows the use of a prior which reflects prior knowledge.
- A Jeffreys' Prior is an improper prior.
- The Wallace-Freeman uses the Fisher Information Matrix term as a statement of the uncertainty with which we will state a parameter's value. The Fisher term is not interpreted as having any association with our prior beliefs. It is entirely sensible that the region of uncertainty will decrease as we sample more data.

5.3.1 An Example of a Jeffreys' Prior

If we consider a Gaussian model to be a function of the mean, μ, and the standard deviation, σ, then the Fisher Information matrix is[10]:

$$F(\mu,\ \sigma) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{2N}{\sigma^2} \end{bmatrix}$$

where N is the number of measurements made. Therefore the Jeffreys' Prior is:

$$Prob(\mu,\ \sigma) \propto \sqrt{det(F(\mu,\ \sigma))} = \frac{\sqrt{2}\ N}{\sigma^2}$$

The Wallace-Freeman region of uncertainty is of volume:

$$UncertaintyVolume(\mu,\ \sigma) \propto \frac{1}{\sqrt{det(F(\mu,\ \sigma))}} = \frac{\sigma^2}{\sqrt{2}\ N}$$

The Jeffreys' Prior interpretation suggests we have small prior belief in large values of σ, while the Wallace-Freeman interpretation suggests we are less willing to precisely state μ and σ if the data either (a) has a large variance, or (b) has a small N.

5.3.2 A Second Example of a Jeffreys' Prior

The determinant of the Fisher Information matrix for the von Mises likelihood function (discussed in Section 3.1) is[11]:

$$det(F(\kappa,\ \mu)) = \kappa\ N^2\ A(\kappa)\ \frac{dA(\kappa)}{d\kappa}$$

where N is the number of measurements made and $A(\kappa)$ is the ratio of two modified Bessel functions:

$$A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$$

Consider the situation as $\kappa \rightarrow 0$. We find that $A(\kappa) \rightarrow 0$ and $\frac{dA(\kappa)}{d\kappa} \rightarrow \frac{1}{2}$. Therefore, $det(F(\kappa,\ \mu)) \rightarrow 0$ as $\kappa \rightarrow 0$.

Adopting a Jeffreys' Prior asserts that we have very little belief that a von Mises distribution will have $\kappa \rightarrow 0$. However, the Wallace-Freeman interpretation suggests that it is difficult to identify parameter values near $\kappa = 0$. We prefer the Wallace-Freeman interpretation, since in situations such as modelling the arrival of patients into a hospital with a von Mises distribution, we do not want to preclude the possibility that patients arrive at a hospital uniformly round the clock.

[10] The Fisher Information matrix for a Gaussian model is derived in Baxter and Oliver [3, Page 20,21].

[11] The Fisher Information matrix for a von Mises model is derived in Wallace and Dowe [30, Page 7].

5.4 Lemma 1 — Invariance of the Wallace-Freeman MML Method

Lemma (given in Wallace [27, Chapter 5]):
Consider a data set, D, with a likelihood $f(D \mid \theta)$ in some convenient parameterisation (of dimension d) with a prior $h(\theta)$ defined in that parameterisation. If we form a new parameterisation (also of dimension d), by applying a differentiable transformation, $g(.)$, such that $\phi = g(\theta)$ then, where the Fisher Information matrices are defined, and their determinants are non-zero[12]:

$$MessLen(\phi \;\&\; D) \;=\; MessLen(\theta \;\&\; D)$$

Proof:
We use Equation (23) to calculate the message length of θ & D in the two parameterisation. Let $Diff$ be the difference in message lengths between the two parameterisations:

$$\begin{aligned} Diff &= MessLen(\phi \;\&\; D) \;-\; MessLen(\theta \;\&\; D) \\ &= -\log(\frac{h(\phi)}{\sqrt{det(F(\phi))}}) \;+\; \log(\frac{h(\theta)}{\sqrt{det(F(\theta))}}) \;-\; L(\phi) \;+\; L(\theta) \end{aligned}$$

By invariance of likelihood functions, we know $L(\phi) = L(\theta)$:

$$Diff \;=\; -\log(\frac{h(\phi)}{\sqrt{det(F(\phi))}}) \;+\; \log(\frac{h(\theta)}{\sqrt{det(F(\theta))}}) \tag{25}$$

Transforming a prior distribution involves dividing it by the Jacobian of the transformation:

$$h(\phi) \;=\; \frac{h(\theta)}{J} \tag{26}$$

where J is:

$$J \;=\; det \begin{vmatrix} \frac{\partial\phi_1}{\partial\theta_1} & \frac{\partial\phi_2}{\partial\theta_1} & \cdots & \frac{\partial\phi_d}{\partial\theta_1} \\ \frac{\partial\phi_1}{\partial\theta_2} & \frac{\partial\phi_2}{\partial\theta_2} & \cdots & \frac{\partial\phi_d}{\partial\theta_2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial\phi_1}{\partial\theta_d} & \frac{\partial\phi_2}{\partial\theta_d} & \cdots & \frac{\partial\phi_d}{\partial\theta_d} \end{vmatrix}$$

Transforming the determinant of the Fisher Information matrix involves dividing it by the square of the Jacobian of the transformation:

$$det(F(\phi)) \;=\; \frac{det(F(\theta))}{J^2} \tag{27}$$

Substituting Equations (26) and (27) into Equation (25) gives:

$$Diff \;=\; -\log(\frac{h(\theta)/J}{\sqrt{det(F(\theta))/J^2}}) \;+\; \log(\frac{h(\theta)}{\sqrt{det(F(\theta))}}) \;=\; 0$$

QED.

5.5 Invariance Examples for the Wallace-Freeman MML Method

5.5.1 A Gaussian Model

In Section 5.3, we gave the determinant of the Fisher Information matrix for a normal model in two parameterisations; in the (mean, standard deviation) frame, $det(F(\mu, \sigma)) = \frac{2\,N^2}{\sigma^4}$ and in the (mean, variance) frame, $det(F(\mu, v)) = \frac{N^2}{2\,v^3}$.

Let us assume a prior distribution over values of σ:

$$h(\sigma) \;=\; \frac{\sigma}{2\pi(1 \;+\; \sigma^2)^{\frac{3}{2}}}$$

[12] The requirement that the determinant of a Fisher Information matrix be non-zero is used here, but is not strictly necessary (for example Wallace and Freeman [32] or Wallace and Dowe [30, Page 8]).

Then the equivalent prior in the (mean, variance) frame is $h(v) = \frac{h(\sigma)}{J}$ where $J = \frac{\partial v}{\partial \sigma} = 2\sigma$, and hence:

$$h(v) = \frac{1}{4\pi(1 + v)^{\frac{3}{2}}}$$

We find the ratio of prior to square root of the Fisher Information remains the same in each parameterisation:

$$\frac{h(\sigma)}{\sqrt{det(F(\mu,\sigma))}} = \frac{h(v)}{\sqrt{det(F(\mu,v))}}$$

$$\frac{\sigma}{2\pi(1 + \sigma^2)^{\frac{3}{2}}} \times \frac{\sigma^2}{\sqrt{2}\ N} = \frac{1}{4\pi(1 + v)^{\frac{3}{2}}} \times \frac{\sqrt{2}\ v^{\frac{3}{2}}}{N}$$

5.5.2 MML Applied to the von Mises Problem

The von Mises problem discussed in Section 3.1 has the property that the mode of the posterior moves when we apply non-linear transformations.

If we apply the Wallace-Freeman MML method to this problem (as was done in Wallace and Dowe [30]), we find that the mode moving is not an obstacle. For the data given in Section 3.1 (sampled from a $\kappa = 0$ distribution) and the prior

$$h_{KM}(\kappa,\ \mu) = \frac{\kappa}{2\pi(1 + \kappa^2)^{\frac{3}{2}}}$$

the Wallace-Freeman MML method gives the estimate ($\kappa = 0.27,\ \mu = 27.8$) (which is independent of co-ordinate system). The mode of the posterior in polar co-ordinates occurs at ($\kappa = 0.53,\ \mu = 27.8$), while in cartesian co-ordinates the mode of the posterior occurs at ($\kappa = 0.25,\ \mu = 27.8$).

Consider an alternative prior:

$$h_{KM}(\kappa,\ \mu) = \frac{1}{\pi^2(1 + \kappa^2)}$$

With this prior the Wallace-Freeman MML estimate is ($\kappa = 0.092,\ \mu = 27.8$). The mode of the posterior in polar co-ordinates occurs at ($\kappa = 0.28,\ \mu = 27.8$). The mode of the posterior suffers from a singularity in cartesian co-ordinates, and hence is maximised at ($\kappa = 0,\ \mu = ?$), independent of the data.

6 Bayesian Model Class Selection

6.1 Definition of Model and Model Class

The term "model" is used inconsistently in the literature. We therefore introduce some terminology to define the problems we are considering. We use the term *model* to mean a model with values given to all its parameters. A *model class* is a set of models with the same parametric form and the same number of parameters. For example, the set of neural nets with 1 hidden layer, and 3 input nodes, 6 hidden nodes, and 2 output nodes would constitute a model class.

6.2 Levels of Inference

In some Bayesian literature, model class selection is considered a distinct task from parameter estimation [16, 13, 22]. Using MacKay's terminology [16], parameter estimation is level one inference and model class selection is level two inference. For example, deciding on the structure of a neural net (i.e., deciding on the number of hidden layers, and the number of units in the hidden layers) is a level two inference, assigning weights to a neural net with a given structure is a level one inference.

6.3 Bayes Factors

Bayes Factors[13] are used for performing level two inference. Consider the problem of selecting between two model classes MC_1 and MC_2 with parameter vectors θ_1 and θ_2 (we can generalize easily to more than two

[13] The term Bayes Factor was first used by Good [9], who attributes the method to Turing, in addition to and independently of Jeffreys [10].

model classes). By Bayes' theorem, the posterior probability that MC_1 is the correct model (assuming one of MC_1 or MC_2 is) is:

$$Prob(MC_1 \mid D) \;=\; \frac{Prob(D \mid MC_1)\; Prob(MC_1)}{Prob(D \mid MC_1)\; Prob(MC_1) \;+\; Prob(D \mid MC_2)\; Prob(MC_2)} \tag{28}$$

The extent to which the data supports MC_1 over MC_2 is measured by the posterior odds for MC_2 against MC_1. By Equation (28):

$$\frac{Prob(MC_2 \mid D)}{Prob(MC_1 \mid D)} \;=\; \frac{Prob(D \mid MC_2)}{Prob(D \mid MC_1)}\frac{Prob(MC_2)}{Prob(MC_1)} \tag{29}$$

where $Prob(D \mid MC_1)$ is the *evidence* for the data given Model Class 1, and $Prob(MC_1)$ is the prior probability we associate with Model Class 1. The *Bayes Factor* is the ratio of the evidences in Equation (29). In many circumstances (for example, neural nets) evaluating the evidence is not an easy calculation. In the next section, we estimate the evidence using Laplace's approximation.

6.4 Approximating the Evidence of a Model Class

We evaluate $Prob(D \mid MC)$ by integrating over all the possible values of the parameter vector which defines MC. Let θ be the parameter vector (of dimension d). Then:

$$Prob(D \mid MC) \;=\; \int Prob(\theta \mid MC) \;\times\; Prob(D \mid \theta)\; d\theta \tag{30}$$

$$Evidence \;=\; \int prior \;\times\; likelihood\; d\theta \tag{31}$$

We term $f(\theta,\; D,\; MC) \;=\; Prob(D \mid \theta)\; Prob(\theta \mid MC)$ as an *un-normalised posterior*. One approach is to use Laplace's method to approximate the integral (e.g., Cheeseman et al. [5] Kass and Raftery [13] MacKay [16] and Raftery [22]). If the un-normalised posterior has a mode at $\hat{\theta}$. We assume that the prior is approximately constant (with value $Prob(\hat{\theta} \mid MC)$) in the region of $\hat{\theta}$:

$$Prob(D \mid MC) \;=\; Prob(\hat{\theta} \mid MC) \int Prob(D \mid \theta)\; d\theta \tag{32}$$

Let I be the integral of the likelihood ($I \;=\; \int Prob(D \mid \theta)\; d\theta$). In Appendix 3, we use Laplace's Method to achieve the following approximation to I [7]:

$$I \;\approx\; Prob(D \mid \hat{\theta})\frac{(2\pi)^{\frac{d}{2}}}{\sqrt{det(M(\hat{\theta}))}} \tag{33}$$

This approximation is equivalent to approximating the likelihood, $Prob(D \mid \theta)$ in the vicinity of the peak, $\hat{\theta}$, with a Gaussian whose covariance matrix is constructed to be the second derivatives of the log-likelihood at the peak. Using this approximation, we find that the evidence is:

$$Prob(D \mid MC) \;\approx\; Prob(\hat{\theta} \mid MC)\; Prob(D \mid \hat{\theta})\frac{(2\pi)^{\frac{d}{2}}}{\sqrt{det(M(\hat{\theta}))}} \tag{34}$$

where $M(\hat{\theta})$ is the observed Fisher Information matrix evaluated at $\hat{\theta}$.

6.5 Laplace's Method Applied to the von Mises Problem

Let us consider the von Mises problem discussed in Sections 3.1 and 5.5.2. If we wish to approximate the evidence of a von Mises distribution using Laplace's approach then we find the mode of the posterior and apply Equation (34) to the mode. However, the parameterisation we select will effect the result. Furthermore, the modes (in the two parameterisations considered) are different from the MML estimate given in Section 5.5.2.

6.6 Comparison with the MML Method

Note that if we take the negative logarithm of Equation (34) then we get an expression similar to Equation (23):

$$\boxed{-\log Prob(D \mid MC) \approx -\frac{d}{2}\, log(2\pi) - log(Prob(\hat{\theta} \mid MC)) + \frac{1}{2} log(det(M(\hat{\theta}))) + L(\hat{\theta}) \quad (35)}$$

MacKay therefore concludes that [15]:

> "With care, therefore, one can replicate Bayesian results in MDL terms. Although some of the earliest work on complex model comparison involved the MDL framework [19], MDL has no apparent advantages over the direct probabilistic approach".

While Equations (34) and (23) are superficially similar, they have different philosophies behind them, are used for different purposes, and have different properties:

- Firstly, the MML equation (23) is used to perform parameter estimation (Level 1 inference); the evidence equation (34) is used to determine the "best" model class (Level 2 inference).
- Secondly, the MML equation (23) approximates the message length of the data using an arbitrary parameter θ; there is no requirement that θ is a mode of the likelihood surface. The evidence equation (34) requires us to have found a mode, $\hat{\theta}$, of the likelihood surface.
- Thirdly, the MML equation (23) is invariant under non-linear parameter transformations; the evidence equation (34) is not invariant.

Barron [2] saw the coding analogy and considered substituting the observed Fisher, $M(\theta)$, for the expected Fisher, $F(\theta)$. The difference between the Wallace-Freeman MML method and Laplace's method is then the lattice constants, as opposed to hyper-sphere constants, and the differences in interpretation.

7 Conclusion

In this paper, we have contrasted MML estimation with

- Bayesian estimation using loss functions.
- Bayesian model selection using Laplace's approximation.

Traditional Bayesian estimation based on the posterior density is un-principled without a loss function being specified. The estimate chosen depends on the loss function; the loss function depends upon the parameterisation. For continuous parameters, we contend that estimation without loss functions is a valid goal of data analysis in some circumstances. It is in these circumstances that the MML estimate is the preferred estimate.

MML estimates the most probable parameter values based on an approximately optimal discretisation of the parameter space. The level of discretisation is determined the average curvature of the likelihood function, the amount of data, and the number of parameters. The discretisation changes under parameter transformations in a compatible way to the manner in which a prior changes, resulting in an invariant estimate of the probability of a model parameter. In addition, this discretisation argument offers a principled method for determining regions of uncertainty for a parameter estimate.

MML appears to be related to other Bayesian methods (notably the adoption of a Jeffreys' Prior and Laplace's method). In the case of the adoption of a Jeffreys' Prior, MML interprets the $\frac{1}{det(Fisher)}$ term in a completely different manner. MML views the Fisher term as corresponding to the uncertainty with which we are willing to state parameter values; the Jeffreys' Prior is difficult to interpret as representing prior beliefs.

In the case of Laplace's method, we find that while the mathematics are similar, the two approaches (MML; Laplace's method) are distinct. The assumptions and pre-conditions are different. We contend that MML is the preferred method here because:

- Laplace's method requires us to identify a preferred parameterisation; MML will give us the same results in any reasonable parameterisation.
- Laplace's method requires we find modes of the posterior, and provide a method for dealing with multimodal posteriors.

8 Acknowledgments

We would like to thank Chris Wallace, David Dowe, Catherine Scipione, Wray Buntine and Lloyd Allison for valuable discussions, and David Hand and Chris Jones for motivating this work.

References

[1] H. Akaike. Factor Analysis and AIC. *Psychometrika*, 52(3):317–332, 1987.

[2] A. Barron. *Logically smooth density estimation.* PhD thesis, Dept. Elect. Eng., Stanford Univ., August 1985.

[3] R.A. Baxter and J.J. Oliver. MDL and MML: Similarities and differences. Technical report TR 207, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1994. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.

[4] J.M. Bernado and A.F.M. Smith. *Bayesian Theory.* Wiley, New York, 1994.

[5] P. Cheeseman, M. Self, J. Kelly, W. Taylor, D. Freeman, and J. Stutz. Bayesian classification. In *Seventh National Conference on Artificial Intelligence*, pages 607–611, Saint Paul, Minnesota, 1988.

[6] J.H. Conway and N.J.A Sloane. *Sphere Packings, Lattices and Groups.* Springer-Verlag, London, 1988.

[7] N.G. de Bruijn. *Asymptotic Methods for Analysis.* North-Holland, Amsterdam, 1970.

[8] N.I. Fisher. *Statistical Analysis of Circular Data.* Cambridge University Press, Cambridge, 1993.

[9] I.J. Good. Explicativity, corroboration, and the relative odds of hypotheses. In *Good thinking— The Foundations of Probability and its applications.* University of Minnesota Press, Minneapolis,MN, 1983.

[10] H. Jeffreys. Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc.*, 31:203–222, 1935.

[11] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. of the Royal Soc. of London A*, 186:453–454, 1946.

[12] H. Jeffreys. *Theory of Probability.* Cambridge, 1961.

[13] R.E. Kass and A.E. Raftery. Bayes Factors and model uncertainty. Technical Report 571, Dept. of Statistics, Carnegie-Mellon University, Dec. 1993.

[14] T.J. Loredo. From Laplace to Supernova SN 1987A: Bayesian inference in astrophysics. In P.F. Fougere, editor, *Maximum Entropy and Bayesian Methods*, pages 81–142. Kluwer Academic Publishers, 1990.

[15] D. J. C. MacKay. Bayesian methods for backpropagation networks. In E. Domany, J. L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks III*, chapter 6. Springer-Verlag, New York, 1994.

[16] David J.C. MacKay. *Bayesian Modeling and Neural Networks.* PhD thesis, Dept. of Computation and Neural Systems, CalTech, 1992.

[17] R.M. Neal. *Bayesian Learning for Neural Networks.* PhD thesis, Graduate Dept. of Computer Science, Univ. of Toronto, March 1995.

[18] J.J. Oliver and D.J. Hand. Introduction to minimum encoding inference. Technical report TR 4-94, Dept. of Statistics, Open University, Walton Hall, Milton Keynes, MK7 6AA, UK, 1994. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.

[19] J.D. Patrick and C.S. Wallace. Stone circle geometries: An information theory approach. In D.C. Heggie, editor, *Archeoastronomy in the Old World.* Cambridge University Press, Cambridge, 1982.

[20] S.J. Press. *Bayesian Statistics.* Wiley, 1989.

[21] S.J. Press and K. Shigemasu. Bayesian inference in factor analysis. In L. Gleser, M. Perlman, S.J. Press, and A. Sampson, editors, *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin.* Springer-Verlag, New York, 1989.

[22] A.E. Raftery. Approximate Bayes Factors and accounting for model uncertainty in generalized linear models. Technical Report 255, Dept. of Statistics, University of Washington, 1993.

[23] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

[24] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society (Series B)*, 49:223–239, 1987.

[25] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore, 1989.

[26] C.S. Wallace. Classification by minimum-message-length inference. In G. Goos and J. Hartmanis, editors, *Advances in Computing and Information – ICCI '90*, pages 72–81. Springer-Verlag, Berlin, 1990.

[27] C.S. Wallace. MML book (in preparation). 1993.

[28] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.

[29] C.S. Wallace and D.M. Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11–34, 1975.

[30] C.S. Wallace and D.L. Dowe. MML estimation of the von Mises concentration parameter. Technical report TR 193, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1993.

[31] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society (Series B)*, 49:240–252, 1987.

[32] C.S. Wallace and P.R. Freeman. Single factor analysis by MML estimation. *Journal of the Royal Statistical Society (Series B)*, 54:195–209, 1992.

[33] C.S. Wallace and J.D. Patrick. Coding decision trees. *Machine Learning*, 11:7–22, 1993.

9 Appendix 1 — Form of the Reparameterization Transformation

To simplify the integral in Equation (20), we required a transformation matrix B which satisfied:

$$\bar{y} \;=\; B^{-1}\bar{x} \quad and \quad \bar{x}^T\; F(\theta)\; \bar{x} \;=\; \bar{y}^T\; \bar{y}$$

B consists of a rotation R followed by a scaling S:

$$B \;=\; R\; S$$

R is a rotation matrix consisting of the eigenvectors of $F(\theta)$:

$$R \;=\; |\bar{e_1} \;\; \bar{e_2} \;\; \ldots \;\; \bar{e_d}|$$

where $\bar{e_i}$ is the i^{th} eigenvector of $F(\theta)$. S is a scaling matrix which takes the form:

$$S \;=\; \begin{vmatrix} \frac{1}{\sqrt{\lambda_1}} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sqrt{\lambda_2}} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \frac{1}{\sqrt{\lambda_d}} \end{vmatrix}$$

where λ_i is the i^{th} eigenvalue of $F(\theta)$.
Substituting $\bar{x} \;=\; B\;\bar{y}$ and hence $(\bar{x}^T \;=\; \bar{y}^T\; B^T)$, we get:

$$\bar{x}^T\; F(\theta)\; \bar{x} \;=\; \bar{y}^T\; B^T\; F(\theta)\; B\; \bar{y}$$

We note:

$$F(\theta)\; B \;=\; F(\theta)\; R\; S \;=\; F(\theta)\; |\bar{e_1} \;\; \bar{e_2} \;\; \ldots \;\; \bar{e_d}|\; S$$

Since $F(\theta)\; \bar{e_i} \;=\; \lambda_i\; \bar{e_i}$, we find

$$F(\theta)\; B \;=\; |\bar{e_1} \;\; \bar{e_2} \;\; \ldots \;\; \bar{e_d}|\; E\; S \;=\; R\; E\; S$$

where

$$E \;=\; \begin{vmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \lambda_d \end{vmatrix}$$

Hence:

$$\bar{x}^T\; F(\theta)\; \bar{x} \;=\; \bar{y}^T\; S^T\; R^T\; R\; E\; S\; \bar{y}$$

As R is a rotation matrix $R^T\; R \;=\; I$, and by the definition of S, $S^T\; E\; S \;=\; I$, and so

$$\bar{x}^T\; F(\theta)\; \bar{x} \;=\; \bar{y}^T\; \bar{y}$$

10 Appendix 2 — Jacobian of the Reparameterization Transformation

The $det(F(\theta))$ can be expressed in terms of its eigenvalues:

$$det(F(\theta)) = \prod_{i=1}^{d} \lambda_i$$

The Jacobian of B^{-1} is then:

$$Jacobian(B^{-1}) = Jacobian(S^{-1})\ Jacobian(R^{-1})$$

Noting that the Jacobian of R^{-1} is 1 (as R^{-1} is a rotation matrix), we get:

$$Jacobian(B^{-1}) = Jacobian(S^{-1}) = \prod_{i=1}^{d} \sqrt{\lambda_i} = \sqrt{\prod_{i=1}^{d} \lambda_i} = \sqrt{det(F(\theta))}$$

11 Appendix 3 — Laplace's Integral

Let $L(\theta)$ be the negative log of the likelihood function:

$$L(\theta) = -\log(Prob(D|\theta)) \tag{36}$$

Consider a Taylor series expansion of $L(\theta)$ about $\bar{\theta}$, the value of θ that maximizes the posterior density (if we assume that the prior $Prob(\theta)$ is constant near the peak of the posterior then this will also be a peak of the likelihood $Prob(D|\theta)$). The expansion is

$$L(\theta) \;=\; L(\bar{\theta}) \;+\; (\theta-\bar{\theta})^T L'(\bar{\theta}) \;-\; \frac{1}{2}(\theta-\bar{\theta})^T M(\bar{\theta})(\theta-\bar{\theta}) \;+\; O(||\theta-\bar{\theta}||^2) \tag{37}$$

where $L'(\theta)$ is the vector of first partial derivatives of $L(\theta)$:

$$L'(\theta) = (\frac{\partial L(\theta)}{\partial \theta_1}, \;\ldots,\; \frac{\partial L(\theta)}{\partial \theta_d})^T \tag{38}$$

d is the dimension of θ, and $M(\theta)$ is observed Fisher Information matrix:

$$M(\theta) \;=\; -\begin{vmatrix} \frac{\partial^2 L(\theta)}{\partial \theta_1^2} & \frac{\partial^2 L(\theta)}{\partial \theta_1 \partial \theta_2} & \ldots & \frac{\partial^2 L(\theta)}{\partial \theta_1 \partial \theta_d} \\ \frac{\partial^2 L(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L(\theta)}{\partial \theta_2 \partial \theta_2} & \ldots & \frac{\partial^2 L(\theta)}{\partial \theta_2 \partial \theta_d} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 L(\theta)}{\partial \theta_d \partial \theta_1} & \frac{\partial^2 L(\theta)}{\partial \theta_d \partial \theta_2} & \ldots & \frac{\partial^2 L(\theta)}{\partial \theta_d \partial \theta_d} \end{vmatrix} \tag{39}$$

At the peak of the posterior, $L'(\bar{\theta}) = 0$. We then define $\bar{L}$ to be the Taylor series expansion of $L(\theta)$ up to second order terms:

$$\bar{L}(\theta) \;=\; L(\bar{\theta}) \;-\; \frac{1}{2}(\theta-\bar{\theta})^T M(\bar{\theta})(\theta-\bar{\theta}) \tag{40}$$

We wish to integrate the likelihood function:

$$I \;=\; \int Prob(D \mid \theta) d\theta \tag{41}$$

$$\;=\; \int e^{\bar{L}(\theta)}\, d\theta \tag{42}$$

$$\;=\; \int e^{L(\bar{\theta})}\, e^{-\,\frac{1}{2}(\theta-\bar{\theta})^T M(\bar{\theta})(\theta-\bar{\theta})}\, d\theta \tag{43}$$

We use Laplace's method for the integral:

$$\int e^{-\,\frac{1}{2}(\theta-\bar{\theta})^T M(\bar{\theta})(\theta-\bar{\theta})}\, d\theta \;=\; \frac{(2\pi)^{\frac{d}{2}}}{\sqrt{det(M(\bar{\theta}))}} \tag{44}$$

Substituting Equation (44) into Equation (43) gives us:

$$I \;=\; e^L(\bar{\theta})\frac{(2\pi)^{\frac{d}{2}}}{\sqrt{det(M(\bar{\theta}))}} \tag{45}$$

$$\;=\; Prob(D \mid \bar{\theta})\frac{(2\pi)^{\frac{d}{2}}}{\sqrt{det(M(\bar{\theta}))}} \tag{46}$$