The Minimum Message Length Principle for Inductive Inference

Daniel F. Schmidt

Centre for Molecular, Environmental, Genetic & Analytic (MEGA) Epidemiology
School of Population Health
University of Melbourne

University of Helsinki, August 25, 2009
Motivation
Coding
MML
MML87
Example
We have observed $n$ data points $y^n = (y_1, \ldots, y_n)$ from some unknown, probabilistic source $p^*$, i.e.

$$y^n \sim p^*$$

where $y^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$.

We wish to learn about $p^*$ from $y^n$.

More precisely, we would like to discover the generating source $p^*$, or at least a good approximation of it, from nothing but $y^n$. 
To approximate $p^*$ we will restrict ourself to a set of potential statistical models.

Informally, a statistical model can be viewed as a conditional probability distribution over the potential dataspace $\mathcal{Y}^n$

$$p(y^n|\theta), \ \theta \in \Theta$$

where $\theta = (\theta_1, \ldots, \theta_k)$ is a parameter vector that indexes the particular model.

Such models satisfy

$$\int_{y^n \in \mathcal{Y}^n} p(y^n|\theta) dy^n = 1$$

for a fixed $\theta$. 
An example would be the univariate normal distribution.

\[ p(y^n|\theta) = \left( \frac{1}{2\pi \tau} \right)^\frac{n}{2} \exp \left( -\frac{1}{2\tau} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \]

where
- \( \theta = (\mu, \tau) \) are the parameters
- \( Y^n = \mathbb{R}^n \)
- \( \Theta = \mathbb{R} \times \mathbb{R}_+ \)
This talk follows the slight abuse of terminology used by Chris Wallace in calling a member of $\Theta$ a *model*

Also referred to as a *fully specified model*

This is because, in the MML framework, there is no real distinction between *structural parameters* that specify a model class and what are traditional termed the *parameter estimates* or *point estimates*
Content

1 Motivation

2 Coding

3 MML

4 MML87

5 Example
MML is based on information theory and coding
Consider a countable set of symbols $\mathcal{X}$ (an alphabet)
Wish to label them by strings of binary digits
⇒ Labelling must be *decodable*
For example, $\mathcal{X} = \{A, C, G, T\}$
- Possible coding, $A = 00$, $C = 01$, $G = 10$, $T = 11$
- or $A = 1$, $C = 01$, $G = 001$, $T = 0001$
- and so on ...

Desire this labelling to be optimal, in some sense
Problem central to compression and information transmission
Assume distribution of symbols given by \( p(x), x \in X \)
Let \( l : X \rightarrow \mathbb{R}_+ \) denote the codelength function
⇒ want our code to be short on average, w.r.t. \( p(\cdot) \)
Restrict ourself to decodable codes ; the solution of

\[
\arg \min_l \left\{ \sum_{x \in X} p(x) l(x) \right\}
\]

is

\[- \log_2 p(x)\]

High probability ⇒ short codeword
Low probability ⇒ long codeword

We use natural log, \( \log \); base \( e \) digits (nits, or nats)
Content

1. Motivation
2. Coding
3. MML
4. MML87
5. Example
Minimum Message Length

- Developed primarily by Chris Wallace with collaborators since 1968
- Connects the notion of compression with statistical inference and model selection
- We frame the problem as one of transmitting the data efficiently from a transmitter to a receiver
  - First, a model from the parameter space $\Theta$ is named by the transmitter (the assertion)
  - Then the data $y^n$ is transmitted to the receiver using this model (the detail)
- For example, in the normal case, the transmitter would name particular values of $(\mu, \tau)$ that can then be used to transmit the data $y^n$
Transmitter and receiver must agree on a common language.

In MML, this is a prior $\pi(\cdot)$ over $\Theta$.

$\Rightarrow$ MML is a Bayesian approach.

The ingredients we need are:

- A model class/family, i.e. linear regression models or neural networks, etc. parameterised by the vector $\theta \in \Theta$.
- A prior probability distribution $\pi(\cdot)$ over $\Theta$.

The receiver only has knowledge of these two things.

But $\Theta$ is uncountable ...
Choose a countable subset $\Theta_* \subset \Theta$

⇒ Discretisation of the parameter space

May now devise a code for members of $\Theta_*$ using $\pi(\cdot)$
Choose a countable subset $\Theta_* \subset \Theta$

$\Rightarrow$ Discretisation of the parameter space

May now devise a code for members of $\Theta_*$ using $\pi(\cdot)$

The transmitter communicates the data to the receiver using a two-part message

- The first part, or assertion, has length $I(\theta)$ and names one model $\theta$ from $\Theta_*$. 
Two-part Messages, Part 1

- Choose a countable subset $\Theta_* \subset \Theta$
  - Discretisation of the parameter space
- May now devise a code for members of $\Theta_*$ using $\pi(\cdot)$
- The transmitter communicates the data to the receiver using a two-part message
  - The first part, or assertion, has length $I(\theta)$ and names one model $\theta$ from $\Theta_*$
  - The second part, or detail, has length $I(y^n|\theta)$, and sends the data $y^n$ using the named model $\theta$
Two-part Messages, Part 1

- Choose a countable subset $\Theta_* \subset \Theta$
  $\implies$ Discretisation of the parameter space
- May now devise a code for members of $\Theta_*$ using $\pi(\cdot)$
- The transmitter communicates the data to the receiver using a two-part message
  - The first part, or assertion, has length $I(\theta)$ and names one model $\theta$ from $\Theta_*$
  - The second part, or detail, has length $I(y^n|\theta)$, and sends the data $y^n$ using the named model $\theta$
This has total (joint) codelength of

\[ I(y^n, \theta) = I(\theta) + I(y^n|\theta) \]

- \( I(\theta) \) measures the ‘complexity’ of the model
- \( I(y^n|\theta) \) measures the fit of the model to the data
  ⇒ So \( I(y^n, \theta) \) trades off model fit against model capability
- Both complexity and fit measured in same units
The Minimum Message Length Principle

To perform estimation one minimises the joint codelength

$$\hat{\theta}_{\text{MML}}(y^n) = \arg \min_{\theta \in \Theta_*} \{ I(\theta) + I(y^n|\theta) \}$$

- The parameter space $\Theta$ can be enlarged to include models of different structure and thus can be used to perform model selection.
The MML estimates $\hat{\theta}_{\text{MML}}(y^n)$ are invariant under one-to-one re-parameterisations of the parameter space $\Theta$. 
Properties

- The MML estimates $\hat{\theta}_{\text{MML}}(y^n)$ are invariant under one-to-one re-parameterisations of the parameter space $\Theta$.
- Unifies the problem of parameter estimation and model selection.
Properties

- The MML estimates $\hat{\theta}_{\text{MML}}(y^n)$ are invariant under one-to-one re-parameterisations of the parameter space $\Theta$.
- Unifies the problem of parameter estimation and model selection.
- The MML principle always works with fully specified models, that is, by quantising the parameter space we may attach probability masses to parameter estimates.
Properties

- The MML estimates $\hat{\theta}_{\text{MML}}(y^n)$ are invariant under one-to-one re-parameterisations of the parameter space $\Theta$
- Unifies the problem of parameter estimation and model selection
- The MML principle always works with *fully specified models*, that is, by quantising the parameter space we may attach probability masses to parameter estimates
- May use joint message length $I(y^n, \theta)$ to assess $\theta$ even if it is not $\hat{\theta}_{\text{MML}}(y^n)$
Properties

- The MML estimates \( \hat{\theta}_{MML}(y^n) \) are invariant under one-to-one re-parameterisations of the parameter space \( \Theta \).
- Unifies the problem of parameter estimation and model selection.
- The MML principle always works with *fully specified models*, that is, by quantising the parameter space we may attach probability masses to parameter estimates.
- May use joint message length \( I(y^n, \theta) \) to assess \( \theta \) even if it is not \( \hat{\theta}_{MML}(y^n) \).
- Difference in message lengths between two models is approximate negative log-posterior odds.
Constructing the Codes

- The Strict Minimum Message Length (SMML) (Wallace & Boulton, 1975) approach constructs a complete two-part codebook designed to minimise expected codeword length given our priors.
- Unfortunately, it is NP-hard and infeasible for all but simplest of problems.
- Fortunately, we are not interested in the codes as much as their lengths.
  - Under certain assumptions, we can approximate these to a high degree.
Choosing \( \Theta^* \) amounts to partitioning of \( \Theta \)

Idea: rather than construct code for all models in \( \Theta^* \), restrict to constructing code only for the model of interest

Let \( \Omega_\theta \) be a neighbourhood of \( \Theta \) near model \( \theta \) of interest

\( \Rightarrow \) Quantisation cell

Make several assumptions

1. The prior density \( \pi(\cdot) \) is slowly varying in \( \Omega_\theta \)
2. The negative log-likelihood function is approximately quadratic in \( \Omega_\theta \)
3. The Fisher information \( |J(\theta)| > 0 \) for all \( \theta \in \Theta \), where
Wallace-Freeman Approximation (MML87), (2)

- Derivation when $\theta \in \Theta \subset \mathbb{R}$
  - $\Omega_\theta = \left\{ \theta \in \Theta : |\theta - \hat{\theta}| \leq w/2 \right\}$ is a symmetric interval of width $w$ centred on $\theta$

- The code-length for the assertion

$$I_{87}(\theta) = -\log \int_{\Omega_\theta} \pi(\theta) d\theta \approx -\log w \pi(\theta)$$

- Assertion length is inversely proportional to prior mass (volume of $\Omega_\theta$)
  - $\Rightarrow$ The smaller $w$, the longer $I_{87}(\theta)$
Wallace-Freeman Approximation (MML87), (3)

- If the named model $\theta$ was stated \textit{exactly}, i.e. $w = 0$, then the detail would be

$$I(y^n|\theta) = -\log p(y^n|\theta)$$

- As $w > 0$, there is an increase in detail length due to imprecisely stating $\theta$

- By Taylor series expansion, code length for the detail

$$-\frac{1}{\int_{\Omega_\theta} \pi(\theta) d\theta} \int_{\Omega_\theta} \pi(\theta) \log p(y^n|\theta) d\theta \approx -\log p(y^n|\theta) + \frac{1}{w} \int_{\Omega_\theta} \tilde{\theta}^2 J(\theta) d\tilde{\theta}$$

where

$$J(\theta) = -\mathbb{E} \left[ \frac{d^2 \log p(y^n|\bar{\theta})}{d\bar{\theta}^2} \right]_{\bar{\theta} = \theta}$$
Wallace-Freeman Approximation (MML87), (4)

- Total codelength of the message

\[ I_{87}(y^n, \theta) = -\log w \pi(\theta) - \log p(y^n|\theta) + \frac{w^2 J(\theta)}{24} \]

- Minimising w.r.t. \( w \) yields

\[ \hat{w} = \left( \frac{12}{J(\theta)} \right)^{1/2} \]

- MML87 codelength for data and model

\[ I_{87}(y^n, \theta) = -\log \pi(\theta) + \frac{1}{2} \log J(\theta) - \frac{1}{2} \log 12 + \frac{1}{2} - \log p(y^n|\theta) \]

Assertion

Detail
Wallace-Freeman Approximation (MML87), (5)

- In multiple dimensions, MML87 codelength for data and model

\[ I_{87}(y^n, \theta) = - \log \pi(\theta) + \frac{1}{2} \log |J(\theta)| + \frac{1}{2} \log \kappa_k + \frac{p}{2} - \log p(y^n | \theta) \]

**Assertion**

where \( \kappa_k \) is the normalised mean-squared quantisation error per parameter of an optimal quantising lattice in \( k \)-dimensions and

\[ J(\theta) = -E \left[ \frac{\partial \log p(y^n | \bar{\theta})}{\partial \bar{\theta} \partial \bar{\theta}'} \bigg|_{\bar{\theta}=\theta} \right] \]

**Detail**

- Useful approximation

\[ \frac{1}{2} (1 + \log \kappa_k) \approx -\frac{k}{2} \log 2\pi + \frac{1}{2} \log k\pi + \psi(1) \]
Wallace-Freeman Approximation (MML87), (6)

- Assertion length $I_{87}(\theta)$ proportional to $|J(\theta)|$
  ⇒ Models with higher Fisher information ‘more complex’

- To perform inference, solve

$$\hat{\theta}_{87}(y^n) = \arg \min_{\theta} \{I_{87}(y^n, \theta)\}$$

- Assigns a probability mass to all models $\theta \in \Theta$
- Valid even if $\Theta$ includes models from different model classes (i.e. model selection)

Daniel F. Schmidt
Minimum Message Length
Wallace-Freeman Approximation (MML87), (7)

- For suitable model classes, $J(\theta) = nJ_1(\theta)$
  - $J_1(\cdot)$ the per sample Fisher information
- Large sample behaviour, $n \to \infty$ as $k$ held constant

$$I_{87}(y^n, \theta) = - \log p(y^n | \theta) + \frac{k}{2} \log n + O(1)$$

$\Rightarrow$ MML87 is asymptotically BIC

- The $O(1)$ term depends on $J_1(\cdot)$, $\pi(\cdot)$ and $k$
- MML87 estimator sequence converges to Maximum Likelihood estimator sequence (under suitable regularity conditions)

- If $k$ grows with $n$, behaviour is very different!
  $\Rightarrow$ MML estimators often consistent even when ML is not
Wallace-Freeman Approximation (MML87), (7)

Theorem

The MML87 estimator is invariant under differentiable, one-to-one reparameterisations of the likelihood function.

Proof: note that the Fisher information transforms as the square of a density.

This property not shared by common Bayes estimators such as posterior mode or posterior mean.
<table>
<thead>
<tr>
<th></th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Motivation</td>
</tr>
<tr>
<td>2</td>
<td>Coding</td>
</tr>
<tr>
<td>3</td>
<td>MML</td>
</tr>
<tr>
<td>4</td>
<td>MML87</td>
</tr>
<tr>
<td>5</td>
<td>Example</td>
</tr>
</tbody>
</table>
Binomial Distribution (1)

- Consider experiment with probability $\theta_\ast$ of yielding a one and probability $(1 - \theta_\ast)$ of yielding a zero.
- Observe $n$ realisations of this experiment, $y^n$, and wish to estimate $\theta_\ast$.
- Negative log likelihood (up to constants)

\[-\log p(y^n|\theta) = -n_1 \log \theta - (n - n_1) \log(1 - \theta)\]

with $n_1 = \sum_{i=1}^{n} y_i$ the number of ones.
- Maximum Likelihood estimate of $\theta_\ast$

$$\hat{\theta}_{ML}(y^n) = \frac{n_1}{n}$$
Choose a uniform prior, $\pi(\theta) \propto 1$

Fisher information

$$J(\theta) = \frac{n}{\theta(1 - \theta)}$$

$\Rightarrow J(\theta) \to \infty$ as $\theta \to 0$ and $\theta \to 1$

MML87 estimator

$$\hat{\theta}_{87}(y^n) = \frac{n_1 + 1/2}{n + 1}$$

‘Regularises’ the ML estimator towards the maximum entropy model ($\theta = 1/2$)

MML87 estimator possesses finite Kullback-Leibler risk, ML estimator does not

$\Rightarrow$ consider case when $n_1 = 0$ or $n_1 = n$
Closer $\theta$ is to boundary, more accurately it must be stated
\[ \Rightarrow \text{Models within same class can be different complexity} \]
Applications/Extensions/Approximations

- Of course, many more applications ...
  - Linear regression models
  - Decision trees/graphs
  - Mixture modelling
  - ARMA models
  - Neural Networks
  - Causal Networks
  - etc...

- Extension of MML87 to hierarchical Bayes models (Makalic & Schmidt, 2009)

- And other approximations when MML87 does not work ...
  - Adaptive coding (Wallace & Boulton 1969)
  - MMLD (Dowe, 1999)
  - MMC_{em} (Makalic, 2007)
  - MML08 (Schmidt, 2008)
References – Theory

References – Applications

- Schmidt, D. & Makalic, E. MML Invariant Linear Regression. *submitted to 22nd Australasian Joint Conference on Artificial Intelligence*, 2009