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

Content

Problem

- We have observed n data points $\mathbf{y}^n = (y_1, \ldots, y_n)$ from some *unknown*, probabilistic source p^*, i.e.

$$\mathbf{y}^n \sim p^*$$

  where $\mathbf{y}^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$.
- We wish to *learn* about p^* from $\mathbf{y}^n$.
- More precisely, we would like to discover the generating source p^*, or at least a *good* approximation of it, from nothing but $\mathbf{y}^n$

Statistical Models

- 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 $\mathcal{Y}^n$

  $$p(\mathbf{y}^n|\boldsymbol{\theta}), \; \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

  where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is a *parameter* vector that indexes the particular model
- Such models satisfy

  $$\int_{\mathbf{y}^n \in \mathcal{Y}^n} p(\mathbf{y}^n|\boldsymbol{\theta}) \mathrm{d}\mathbf{y}^n = 1$$

  for a *fixed* $\boldsymbol{\theta}$

Statistical Models ...

- An example would be the univariate normal distribution.

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

  where
  - $\boldsymbol{\theta} = (\mu, \tau)$ are the parameters
  - $\mathcal{Y}^n = \mathbb{R}^n$
  - $\boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}_+$

Terminology

- This talk follows the slight abuse of terminology used by Chris Wallace in calling a member of Θ 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

Codebooks

- 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
  $\Rightarrow$ 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

Codebooks

- Assume distribution of symbols given by $p(x)$, $x \in \mathcal{X}$
- Let $l : \mathcal{X} \to \mathbb{R}_+$ denote the codelength function
  $\Rightarrow$ want our code to be short on average, w.r.t. $p(\cdot)$
- Restrict ourself to decodable codes ; the solution of

$$\arg\min_l \left\{ \sum_{x \in \mathcal{X}} p(x) l(x) \right\}$$

  is

$$-\log_2 p(x)$$

  - High probability $\Rightarrow$ short codeword
  - Low probability $\Rightarrow$ long codeword
- We use natural log, $\log$; base e digits (nits, or nats)

Content

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 reciever
  - First, a model from the parameter space Θ is named by the transmitter (the assertion)
  - Then the data $\mathbf{y}^n$ is transmitted to the reciever using this model (the detail)
- For example, in the normal case, the transmitter would name particular values of (μ, τ) that can then be used to transmit the data $\mathbf{y}^n$

Minimum Message Length

- Transmitter and receiver must agree on a common language
- In MML, this is a prior $\pi(\cdot)$ over Θ
  $\Rightarrow$ 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 $\boldsymbol{\theta} \in \Theta$
  - A prior probability distribution $\pi(\cdot)$ over Θ
- The reciever only has knowledge of these two things
- But Θ is uncountable ...

Two-part Messages, Part 1

- Choose a countable subset $\Theta_* \subset \Theta$
  $\Rightarrow$ Discretisation of the parameter space
- May now devise a code for members of Θ_* using $\pi(\cdot)$

Two-part Messages, Part 1

- Choose a countable subset $\Theta_* \subset \Theta$
  $\Rightarrow$ Discretisation of the parameter space
- May now devise a code for members of Θ_* using $\pi(\cdot)$
- The transmitter communicates the data to the reciever using a two-part message
  - The first part, or assertion, has length $I(\boldsymbol{\theta})$ and names one model $\boldsymbol{\theta}$ from Θ_*

Two-part Messages, Part 1

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

Two-part Messages, Part 1

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

Two-part Messages, Part 2

- This has total (joint) codelength of

$$I(\mathbf{y}^n, \boldsymbol{\theta}) = I(\boldsymbol{\theta}) + I(\mathbf{y}^n|\boldsymbol{\theta})$$

- $I(\boldsymbol{\theta})$ measures the 'complexity' of the model
- $I(\mathbf{y}^n|\boldsymbol{\theta})$ measures the fit of the model to the data
  $\Rightarrow$ So $I(\mathbf{y}^n, \boldsymbol{\theta})$ trades off model fit against model capability
- Both complexity and fit measured in same units

Two-part Messages, Part 3

The Minimum Message Length Principle

To perform estimation one minimises the joint codelength

$$\hat{\boldsymbol{\theta}}_{\mathrm{MML}}(\mathbf{y}^n) = \arg\min_{\boldsymbol{\theta} \in \Theta_*} \left\{ I(\boldsymbol{\theta}) + I(\mathbf{y}^n | \boldsymbol{\theta}) \right\}$$

- The parameter space Θ can be enlarged to include models of different structure and thus can be used to perform *model selection*

Properties

- The MML estimates $\hat{\boldsymbol{\theta}}_{\mathrm{MML}}(\mathbf{y}^n)$ are invariant under one-to-one re-parameterisations of the parameter space Θ

Properties

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

Properties

- The MML estimates $\hat{\boldsymbol{\theta}}_{\mathrm{MML}}(\mathbf{y}^n)$ are invariant under one-to-one re-parameterisations of the parameter space Θ
- 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 $\hat{\boldsymbol{\theta}}_{\mathrm{MML}}(\mathbf{y}^n)$ are invariant under one-to-one re-parameterisations of the parameter space Θ
- 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 $I(\mathbf{y}^n, \boldsymbol{\theta})$ to assess $\boldsymbol{\theta}$ even if it is not $\hat{\boldsymbol{\theta}}_{\mathrm{MML}}(\mathbf{y}^n)$

Properties

- The MML estimates $\hat{\boldsymbol{\theta}}_{\mathrm{MML}}(\mathbf{y}^n)$ are invariant under one-to-one re-parameterisations of the parameter space Θ
- 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 $I(\mathbf{y}^n, \boldsymbol{\theta})$ to assess $\boldsymbol{\theta}$ even if it is not $\hat{\boldsymbol{\theta}}_{\mathrm{MML}}(\mathbf{y}^n)$
- 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 codelength given our priors
- Unfortunately, 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

Content

Wallace-Freeman Approximation (MML87), (1)

- Choosing Θ_* amounts to partitioning of Θ
- Idea: rather than construct code for all models in Θ_*, restrict to constructing code only for the model of interest
- Let $\Omega_{\boldsymbol{\theta}}$ be a neighbourhood of Θ near model $\boldsymbol{\theta}$ of interest
  $\Rightarrow$ Quantisation cell
- Make several assumptions
  1. The prior density $\pi(\cdot)$ is slowly varying in $\Omega_{\boldsymbol{\theta}}$
  2. The negative log-likelihood function is approximately quadratic in $\Omega_{\boldsymbol{\theta}}$
  3. The Fisher information $|\mathbf{J}(\boldsymbol{\theta})| > 0$ for all $\boldsymbol{\theta} \in \Theta$, 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 θ
- The codelength for the assertion

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

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

Wallace-Freeman Approximation (MML87), (3)

- If the named model θ was stated *exactly*, i.e. $w = 0$, then the detail would be

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

- As $w > 0$, there is an increase in detail length due to imprecisely stating θ
- By Taylor series expansion, codelength for the detail

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

where

$$J(\theta) = -\mathrm{E}\left[\left. \frac{d^2 \log p(\mathbf{y}^n|\bar{\theta})}{d\bar{\theta}^2} \right|_{\bar{\theta}=\theta} \right]$$

Wallace-Freeman Approximation (MML87), (4)

- Total codelength of the message

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

- Minimising w.r.t. w yields

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

- MML87 codelength for data and model

$$I_{87}(\mathbf{y}^n, \theta) = \underbrace{-\log \pi(\theta) + \frac{1}{2}\log J(\theta) - \frac{1}{2}\log 12}_{Assertion} + \underbrace{\frac{1}{2} - \log p(\mathbf{y}^n|\theta)}_{Detail}$$

Wallace-Freeman Approximation (MML87), (5)

- In multiple dimensions, MML87 codelength for data and model

$$I_{87}(\mathbf{y}^n, \boldsymbol{\theta}) = \underbrace{-\log \pi(\boldsymbol{\theta}) + \frac{1}{2} \log |J(\boldsymbol{\theta})| + \frac{1}{2} \log \kappa_k}_{Assertion} + \underbrace{\frac{p}{2} - \log p(\mathbf{y}^n | \boldsymbol{\theta})}_{Detail}$$

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

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

- 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 $I_{87}(\boldsymbol{\theta})$ proportional to $|\mathbf{J}(\boldsymbol{\theta})|$
  $\Rightarrow$ Models with higher Fisher information ‘more complex’
- To perform inference, solve

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

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

Wallace-Freeman Approximation (MML87), (7)

- For suitable model classes, $\mathbf{J}(\boldsymbol{\theta}) = n\mathbf{J}_1(\boldsymbol{\theta})$
  - $\mathbf{J}_1(\cdot)$ the per sample Fisher information
- Large sample behaviour, $n \to \infty$ as k held constant

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

$\Rightarrow$ MML87 is asymptotically BIC

  - The $O(1)$ term depends on $\mathbf{J}_1(\cdot)$, $\pi(\cdot)$ 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

Content

Binomial Distribution (1)

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

$$-\log p(\mathbf{y}^n|\theta) = -n_1 \log \theta - (n - n_1) \log(1 - \theta)$$

with $n_1 = \sum_{i=1}^n y_i$ the number of ones

- Maximum Likelihood estimate of θ_*

$$\hat{\theta}_{\mathrm{ML}}(\mathbf{y}^n) = \frac{n_1}{n}$$

Binomial Distribution (2)

- Choose a uniform prior, $\pi(\theta) \propto 1$
- Fisher information

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

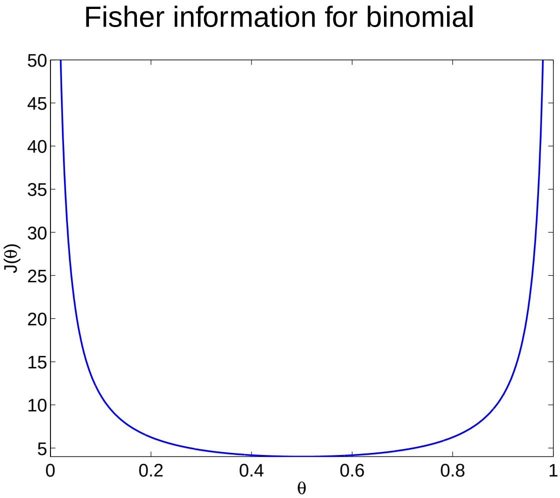
$\Rightarrow J(\theta) \to \infty$ as $\theta \to 0$ and $\theta \to 1$

- MML87 estimator

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

- '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 $n_1 = n$

Binomial Distribution (3)



- Closer θ is to boundary, more accurately it must be stated
  $\Rightarrow$ 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

- Wallace, C. S. & Boulton, D. M. An information measure for classification. *Computer Journal*, 1968, 11, pp. 185–194
- Wallace, C. S. & Boulton, D. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 1975, 3, pp. 11–34
- Wallace, C. S. & Freeman, P. R. Estimation and inference by compact coding. *Journal of the Royal Statistical Society (Series B)*, 1987, 49, pp. 240–252
- Wallace, C. S. False Oracles and SMML Estimators. *Proc. Int. Conf. on Information, Statistics and Induction in Science (ISIS)*, 1996, pp. 304–316
- Wallace, C. S. & Dowe, D. L. Minimum Message Length and Kolmogorov Complexity. *Computer Journal*, 1999, 42, pp. 270–283
- Wallace, C. S. & Dowe, D. L. Refinements of MDL and MML Coding. *Computer Journal*, 1999, 42, pp. 330–337
- Farr, G. E. & Wallace, C. S. The complexity of Strict Minimum Message Length inference. *Computer Journal*, 2002, 45, pp. 285–292
- Wallace, C. S. **Statistical and Inductive Inference by Minimum Message Length.** *Springer*, 2005

References – Applications

- Wallace, C. S. & Freeman, P. R. Single-Factor Analysis by Minimum Message Length Estimation. *Journal of the Royal Statistical Society (Series B)*, 1992, 54, pp. 195–209
- Wallace, C. S. & Patrick, J. D. Coding Decision Trees. *Machine Learning*, 1993, 11, pp. 7–22
- Wallace, C. S. Multiple factor analysis by minimum message length estimation. *Proceedings of the Fourteenth Biennial Statistical Conference*, 1998, pp. 144
- Wallace, C. S. & Korb, K. B. Learning linear causal models by MML sampling. *Causal Models and Intelligent Data Management*, Springer-Verlag, 1999, pp. 89-111
- Wallace, C. S. & Dowe, D. L. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 2000, 10, pp. 73–83
- Schmidt, D. & Makalic, E. MML Invariant Linear Regression. *submitted to 22nd Australasian Joint Conference on Artificial Intelligence*, 2009
- Makalic, E. & Schmidt, D. F. Minimum Message Length Shrinkage Estimation. *Statistics & Probability Letters*, 2009, 79, pp. 1155–1161
- Schmidt, D. F. & Makalic, E. Shrinkage and Denoising by Minimum Message Length. *submitted to IEEE Transactions on Information Theory*, 2009
- Makalic, E. & Schmidt, D. F. Minimum Message Length, Bayesian inference and Nuisance Parameters. *submitted to Biometrika*, 2009