A New Two-Part Codelength Formula For Parameter Estimation and Model Selection

Daniel F. Schmidt

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

WITMSE 2011, Helsinki, August 10, 2011

Content

Some notation

- Let $\mathbf{y}^n = (y_1, \ldots, y_n)' \in \mathcal{Y}^n \subseteq \mathbb{R}^n$ denote observed data
- Let $\gamma \in \Gamma$ denote a model from countable set of models Γ
- Let $p(\mathbf{y}^n|\boldsymbol{\theta})$ denote a set of distributions in model γ indexed by parameter vector $\boldsymbol{\theta} \in \Theta_\gamma$
- Given $\mathbf{y}^n$, wish to select a *plausible* fully specified model that explains the data, i.e., the pair $\left\{\hat{\boldsymbol{\theta}}, \hat{\gamma}\right\}$
- Present some work based on *two-part* codes (Rissanen 1978, Wallace & Freeman 1987)

Two-Part Codes 1

- Examine minimum message length (MML) two-part codes
- Assume a prior $\pi_\gamma(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta_\gamma$ exists for all $\gamma \in \Gamma$
- Assertion : First part of the code, names a $\boldsymbol{\theta} \in \Theta_\gamma$
  - Denote length by $I(\boldsymbol{\theta}; \gamma)$
- Detail : Second part of the code, names a $\mathbf{y}^n \in \mathcal{Y}^n$ using $\boldsymbol{\theta}$ named in assertion
  - Denote length by $I(\mathbf{y}^n | \boldsymbol{\theta}; \gamma)$
- Also need a codeword to name a $\gamma \in \Gamma$, with length $I(\gamma)$
- Small redundancy over one-part codes
  $\Rightarrow$ Unify point estimation *and* model selection

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

Two-Part Codes 2

- Strict MML minimises expected codelengths
  - Minimisation w.r.t. to marginal distribution
  - In general, NP-hard problem
- Under suitable conditions, Wallace–Freeman (MML87) approximate codelength one way to get codelengths

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

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

  - k denotes number of continuous parameters
  - $\mathbf{J}_\gamma(\boldsymbol{\theta})$ is Fisher information matrix
  - κ_k is normalized second moment of the optimal quantising lattice in k dimensions
- Studied by J. Takeuchi as MDL estimator, among others

Two-Part Codes 3

- Quadratic approximation computationally tractable
- Accuracy depends on behaviour of likelihood and prior
- Problems can occur when
  - Fisher information matrix near singular
  - Curvature of prior is large relative to curvature of likelihood
- Desirable to have more robust two-part codelength formula
  $\Rightarrow$ We present one possible formula

Content

Random Coding Model Selection 1

- Wish to transmit data $\mathbf{y}^n$ to receiver using distributions $p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})$ from model γ
- Want to avoid full quantisation of parameter space Θ_γ
- We can do this with codelengths that are *on average* short, using Wallace's ingenious random coding procedure
- Is the basis for the MMLD codelength
  - More robust than quadratic approximations

Random Coding Model Selection 2

- Assume both receiver and transmitter have access to pseudo-random number generator capable of simulating from prior $\pi_\gamma(\boldsymbol{\theta})$
  - Both start with the same seed
- Transmitter draws distributions $\boldsymbol{\theta}_1, \ldots$ from $\pi_\gamma(\cdot)$ until they generate one that lies in some set $S \subseteq \Theta_\gamma$
- Transmitter sends number of draws, say d, to obtain this distribution, say $\boldsymbol{\theta}_d$, to receiver using universal code for integers
  - Denote length of this codeword by $l^*(d)$
- Receiver may repeat process to arrive at $\boldsymbol{\theta}_d$
  $\Rightarrow$ Transmitter may send data $\mathbf{y}^n$ using $p_\gamma(\cdot|\boldsymbol{\theta}_d)$
- Total codelength

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

Random Coding Model Selection 3

- Desire messages to be short on average

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

  Expectation is taken w.r.t. random variables $(d, \boldsymbol{\theta}_d)$

- Let

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

- d is distributed as per a geometric with parameter $q_\gamma(S)$
- From properties of geometric, and using log-star code

$$\begin{aligned} I(S;\gamma) &= \mathrm{E} \left[l^*(d) \right] \\ &= -\log q_\gamma(S) + O(\log \log q_\gamma(S)) \end{aligned}$$

  This is our “assertion”

Random Coding Model Selection 4

- Note that d and $\boldsymbol{\theta}_d$ are independent

$$\begin{aligned} I(\mathbf{y}^n|S;\gamma) &= \mathrm{E}\left[-\log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}_d)\right] \\ &= -\frac{1}{q_\gamma(S)}\int_S \pi_\gamma(\boldsymbol{\theta})\log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})d\boldsymbol{\theta}. \end{aligned}$$

  This is our "detail"

- Let

$$\Omega(\mathbf{y}^n) = \arg\min_{S\subseteq\Theta_\gamma}\left\{I(S;\gamma) + I(\mathbf{y}^n|S;\gamma)\right\}$$

  denote the *uncertainty region*.

Random Coding Model Selection 5

- MMLD codelength

$$I_D(\mathbf{y}^n;\gamma) = I(\gamma) + I(\Omega(\mathbf{y}^n);\gamma) + I(\mathbf{y}^n|\Omega(\mathbf{y}^n);\gamma)$$

- Advantages
  - More robust to assumptions than Wallace–Freeman codelength formula
  - More computationally friendly than SMML
- Not really two-part
- Assertion is set of plausible distributions for observed data
- Does not give measure of quality-of-fit for arbitrary $\boldsymbol{\theta} \in \Theta_\gamma$
  $\Rightarrow$ Cannot be used to explicitly estimate parameters

Content

MML08 Codelength Formula

- Modify MMLD to retain robustness but obtain explicit measure for point estimation
- MMLD "round-off" function

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

  $\Rightarrow$ Excess in detail length due to quantisation of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$
- Depends on the observed data $\mathbf{y}^n$
- Want a codelength function independent of the observed data for any distribution $\boldsymbol{\theta}^* \in \Theta_\gamma$ under consideration

MML08 Codelength Formula 1

- As in Wallace–Freeman derivation assume $\mathbf{y}^n \sim p_\gamma(\cdot|\boldsymbol{\theta}^*)$
  - Treat $\boldsymbol{\theta}^*$ as representative parameter vector in a quantisation cell
- Yields *expected round-off*

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

where

$$\Delta_\gamma(\boldsymbol{\theta}^* || \boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}^*} \left[\log \left(\frac{p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}^*)}{p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})} \right) \right], \tag{3}$$

is the directed Kullback–Leibler divergence for n samples.

MML08 Codelength Formula 2

- Now can define the MML08 "codelength"

$$I_{08}(\mathbf{y}^n, \boldsymbol{\theta}^*; \gamma) = -\log p_\gamma(\mathbf{y}^n|\boldsymbol{\theta}^*) - \log q_\gamma(\Omega_\gamma(\boldsymbol{\theta}^*)) + \frac{1}{q_\gamma(\Omega_\gamma(\boldsymbol{\theta}^*))} \int_{\Omega_\gamma(\boldsymbol{\theta}^*)} \pi_\gamma(\boldsymbol{\theta}) \Delta_\gamma(\boldsymbol{\theta}^*||\boldsymbol{\theta}) d\boldsymbol{\theta}$$

  where $\Omega_\gamma(\boldsymbol{\theta}^*)$ is called the *expected uncertainty* region, and is chosen to minimise $I_{08}(\mathbf{y}^n, \boldsymbol{\theta}^*)$.
- Trade-off determines quantisation
- Model selection, and parameter estimation

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

  $\Rightarrow$ Robust alternative to MML87

Some properties

- **Property 1** : The MML08 codelength is invariant under one-to-one differentiable reparameterisations, assuming suitable transformation of the prior density $\pi_\gamma(\boldsymbol{\theta})$.
  - Ensures that point estimates are also invariant ; not shared by many Bayes estimators
- **Property 2** : The expected uncertainty region satisfies

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

  where $\delta_\gamma(\boldsymbol{\theta}^*)$ is the Kullback–Leibler divergence of any distribution on the boundary of $\Omega_\gamma(\boldsymbol{\theta}^*)$.
  - Uncertainty region is a KL ball ; suggests interesting relationships with distinguishable distributions

Large Sample Behaviour 1

- Under suitable regularity conditions

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

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

- Strictly shorter than MML87 because ellipses do not tile

Large Sample Behaviour, Implications

- Following (Wallace 2005, pp. 238)

$$I_{08}(\mathbf{y}^n, \boldsymbol{\theta}^*; \gamma) = -\log \int_{\Theta_\gamma} p_\gamma(\mathbf{y}^n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} + O(\log k)$$

  Potentially computationally simpler alternative to marginal likelihood/Bayes mixture codes

- Under Jeffrey's prior

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

  $\Rightarrow$ "Almost" minimax regret

Possible Applications

- Autoregressive Moving-Average Models
  - Problematic in pole-space
- MLP Neural Networks
- Hyper-parameter estimation
  - MML/MDL ℓ_1 shrinkage
- Codelengths of cutpoints
  - Change point estimation
  - Bin positions in variable width histograms
- Developing simple MCMC algorithm to compute MML08 codelengths

Possible Applications

- Thank you. Questions ?

References

- 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
- Rissanen, J. J. Modeling by shortest data description. *Automatica*, 14, pp 465–471, 1978
- 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
- Rissanen, J. J. Fisher Information and Stochastic Complexity. *IEEE Transactions on Information Theory*, 1996, 42, pp. 40–47
- Takeuchi, J. Characterization of the Bayes Estimator and MDL Estimator for Exponential Families, 1997, 43, pp. 1165–1174
- 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
- Schmidt, D. F. Minimum Message Length Inference of Autoregressive Moving Average Models. *PhD Thesis*, 2011