

# MML Invariant Linear Regression

Daniel F. Schmidt and Enes Makalic

The University of Melbourne  
Centre for MEGA Epidemiology  
Carlton VIC 3053, Australia  
{dschmidt, emakalic}@unimelb.edu.au

**Abstract.** This paper derives two new information theoretic linear regression criteria based on the minimum message length principle. Both criteria are invariant to full rank affine transformations of the design matrix and yield estimates that are minimax with respect to squared error loss. The new criteria are compared against state of the art information theoretic model selection criteria on both real and synthetic data and show good performance in all cases.

## 1 Introduction

Consider the linear regression model for explaining data  $\mathbf{y}^n \in \mathbb{R}^n$

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where  $\boldsymbol{\beta} \in \mathbb{R}^p$  is the vector of linear regression coefficients,  $\boldsymbol{\varepsilon} \in \mathbb{R}^n$  are i.i.d. variates distributed as per  $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \tau \mathbf{I}_n)$  (where  $\mathbf{I}_k$  denotes the  $(k \times k)$  identity matrix),  $\mathbf{X}_\gamma$  is the design matrix of regressors and  $\gamma \subset \{1, \dots, q\}$  is an index vector determining which regressors comprise the design matrix. Let  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_q)$  be the complete matrix of regressors, where  $\mathbf{x}_i \in \mathbb{R}^n$  and  $q$  is the maximum number of candidate regressors. Given model structure index  $\gamma$ , the model design matrix is defined as

$$\mathbf{X}_\gamma = (\mathbf{x}_{\gamma_1}, \dots, \mathbf{x}_{\gamma_p})$$

Denote the full vector of continuous parameters by  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau) \in \Theta \subset \mathbb{R}^{p+1}$  where  $\Theta$  is the parameter space. The parameters  $\boldsymbol{\theta}$  are considered unknown and must be inferred from the data  $\mathbf{y}^n$ , along with the optimal regressor subset  $\gamma$ .

This paper considers information theoretic model selection criteria based on Minimum Message Length (MML) [1] for inference of linear regression models. The criteria derived here are: (1) invariant under all full rank affine transformations of the design matrix (2) yield estimates that are minimax with respect to the squared error loss and, (3) require no user specified parameters. Most previous MML criteria for linear regression [1–3] are based on ridge regression style priors and none possess all of these attractive properties. In addition, one of the new criteria allows for shrinkage of parameters and selection of relevant regressors to be performed within a single framework. The resultant criteria are closely related to the linear regression criterion derived using the Normalized Maximum Likelihood (NML) [4] universal model.

## 2 Minimum Message Length (MML)

Within the MML framework [1, 5, 6] of inductive inference, the model that best compresses the data resulting in the shortest message length is deemed optimal. The hypothetical message consists of two parts: the assertion,  $I_{87}(\boldsymbol{\theta})$ , which is a statement of the chosen model, and the detail,  $I_{87}(\mathbf{y}^n|\boldsymbol{\theta})$ , which denotes the encoding of the data under the assumption that the model named in the assertion is optimal. The Wallace-Freeman, or MML87 approximation [6], states that a model  $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$  and data  $\mathbf{y}^n = (y_1, \dots, y_n)$  may be concisely transmitted with a message approximately given by

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

where  $\pi(\cdot)$  denotes a prior distribution over the parameter space  $\boldsymbol{\Theta}$ ,  $\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$  is the Fisher information matrix, and  $\kappa_k$  is the normalised second moment of an optimal quantising lattice in  $k$ -dimensions. In this paper, the need to determine  $\kappa_k$  for arbitrary dimension  $k$  is circumvented by using the approximation (pp. 237, [1])

$$\frac{k}{2} (\log \kappa_k + 1) \approx -\frac{k}{2} \log(2\pi) + \frac{1}{2} \log(k\pi) + \psi(1)$$

where  $\psi(\cdot)$  is the digamma function. We define  $\log$  as the natural logarithm, and as such, all message lengths are measured in *nits* (nats), or base- $e$  digits. The MML principle advocates choosing the model  $\hat{\boldsymbol{\theta}}_{87}(\mathbf{y}^n)$  that minimises (1) as the most *a posteriori* likely explanation of the data in the light of the chosen priors.

The original Wallace-Freeman approximation requires that the prior be completely specified. Recently, this requirement has been relaxed by the introduction of a Wallace-Freeman like extension to hierarchical Bayes models in which the parameters and hyperparameters are jointly estimated from the data [7]. If  $\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{\alpha})$  is the prior density over  $\boldsymbol{\theta}$  parametrised by hyperparameters  $\boldsymbol{\alpha}$ , and  $\pi_{\boldsymbol{\alpha}}(\cdot)$  is the prior density over  $\boldsymbol{\alpha}$ , the joint message length of  $\mathbf{y}^n$ ,  $\boldsymbol{\theta}$  and  $\boldsymbol{\alpha}$  is

$$I_{87}(\mathbf{y}^n, \boldsymbol{\theta}, \boldsymbol{\alpha}) = -\log p(\mathbf{y}^n|\boldsymbol{\theta}) - \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha})\pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) + \frac{1}{2} \log |\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})||\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| + \text{const} \quad (2)$$

where

$$\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \mathbb{E} \left[ \frac{\partial^2 I_{87}(\mathbf{y}^n, \hat{\boldsymbol{\theta}}_{87}(\mathbf{y}^n|\boldsymbol{\alpha})|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \right]$$

denotes the Fisher information for the hyperparameters  $\boldsymbol{\alpha}$ , the expectations taken with respect to the marginal distribution  $r(\mathbf{y}^n|\boldsymbol{\alpha}) = \int p(\mathbf{y}^n|\boldsymbol{\theta})\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta}$ .

### 3 Linear Regression with a Uniform Prior

Dropping  $\gamma$  for brevity, the negative log-likelihood function for a linear regression model given a set of parameters  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau) \in \mathbb{R}^{p+1}$  is

$$-\log p(\mathbf{y}^n | \boldsymbol{\theta}) = \frac{n}{2} \log(2\pi) + \frac{n}{2} \log \tau + \frac{1}{2\tau} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (3)$$

The Fisher information for the linear regression model is known to be

$$|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| = \left( \frac{1}{2\tau^{p+2}} \right) |\mathbf{X}'\mathbf{X}| \quad (4)$$

To apply the MML87 formula (1) we require a suitable prior distribution  $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\tau)$  for the regression parameters  $\boldsymbol{\beta}$  and the noise variance  $\tau$ . One aim of the paper is to derive model selection criteria that do not require specification of subjective priors which is often difficult in practice in the linear regression setting. Ideally, one wishes to give each set of feasible regression coefficients  $\boldsymbol{\beta} \in \mathbb{R}^p$  an equal prior probability. A possible method is to use the uniform prior over each coefficient, which is of course improper and requires a bounding of the parameter space to avoid the Jeffreys-Lindley paradox. The data  $\mathbf{y}^n$  are assumed to be generated from the model

$$\mathbf{y} = \mathbf{y}_* + \boldsymbol{\varepsilon}$$

where  $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \tau \mathbf{I}_n)$  and  $\mathbf{y}_*$  is the ‘true’ underlying regression curve. Noting that  $E[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}] = n\tau$  and  $E[\boldsymbol{\varepsilon}'\mathbf{y}] = 0$ , it is clear that

$$E[\mathbf{y}'\mathbf{y}] = \mathbf{y}_*' \mathbf{y}_* + n\tau \quad (5)$$

For a given  $\boldsymbol{\beta}$ , one can construct an estimate of  $\mathbf{y}_*$ , say  $\mathbf{X}\boldsymbol{\beta}$ ; since  $\tau$  is unknown and strictly positive, by (5), we expect this estimate to satisfy

$$\mathbf{y}'\mathbf{y} \geq (\mathbf{X}\boldsymbol{\beta})' (\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}' (\mathbf{X}'\mathbf{X}) \boldsymbol{\beta} \quad (6)$$

that is, we do not expect the estimate of  $\mathbf{y}_*$  to have greater energy than the energy of the observed data  $\mathbf{y}$ . From (6), the feasible parameter hyper-ellipsoid  $A \subset \mathbb{R}^p$  is then given by

$$A = \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\beta}' (\mathbf{X}'\mathbf{X}) \boldsymbol{\beta} \leq \mathbf{y}'\mathbf{y}\}$$

A suitable prior for  $\boldsymbol{\beta}$  is then a uniform prior over the feasible parameter set  $A$

$$\pi(\boldsymbol{\beta}) = \frac{1}{\text{vol}(A)} = \frac{\Gamma(p/2 + 1) \sqrt{|\mathbf{X}'\mathbf{X}|}}{(\pi \mathbf{y}'\mathbf{y})^{(p/2)}}, \boldsymbol{\beta} \in A \quad (7)$$

The prior over  $\tau$  is chosen to be the standard conjugate prior

$$\pi(\tau) \propto \tau^{-\nu} \quad (8)$$

where  $\nu$  is a suitably chosen hyperparameter. The impropriety of (8) is not problematic as  $\tau$  is a common parameter for all regression models. Using (3), (7), (8) and (4) in (1) and minimising the resulting codelength yields

$$\hat{\beta}_{87}(\mathbf{y}^n) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \hat{\beta}_{\text{LS}}(\mathbf{y}^n) \quad (9)$$

$$\hat{\tau}_{87}(\mathbf{y}^n) = \frac{\mathbf{y}'\mathbf{y} - \xi(\mathbf{y}^n)}{n - p + 2\nu - 2} \quad (10)$$

where  $\xi(\mathbf{y}^n) = \hat{\beta}_{87}(\mathbf{y}^n)' (\mathbf{X}'\mathbf{X}) \hat{\beta}_{87}(\mathbf{y}^n)$  is the fitted sum of squares, and  $\hat{\beta}_{\text{LS}}(\mathbf{y}^n)$  are the usual least squares estimates. The final complete message length, up to a constant, evaluated at the MML estimates (9) and (10) is

$$\begin{aligned} I_u(\mathbf{y}^n|\gamma) = & \left(\frac{n-p}{2}\right) \log 2\pi + \left(\frac{n-p+2\nu-2}{2}\right) (\log \hat{\tau}_{87}(\mathbf{y}^n) + 1) + \frac{p}{2} \log(\pi \mathbf{y}'\mathbf{y}) \\ & - \log \Gamma\left(\frac{p}{2} + 1\right) + \frac{1}{2} \log(p+1) + \text{const} \end{aligned} \quad (11)$$

where  $I_u(\mathbf{y}^n|\gamma) \equiv I_u(\mathbf{y}^n, \hat{\beta}_{87}(\mathbf{y}^n), \hat{\tau}_{87}(\mathbf{y}^n)|\gamma)$ . The code (11) is henceforth referred to as the  $\text{MML}_u$  code. We note that as (11) depends on  $\mathbf{X}$  only through  $\hat{\tau}_{87}(\mathbf{y}^n)$ , the message length is invariant under all full rank affine transformations of  $\mathbf{X}$ . We also note that as the MML87 estimates  $\hat{\beta}_{87}(\mathbf{y}^n)$  under the uniform prior coincide with the least-squares estimates they are minimax with respect to squared error loss for all  $p > 0$ .

The criterion (11) handles the case  $p = 0$  (i.e., no signal) gracefully, and is of the same computational complexity as well known asymptotic information criteria such as Akaike's Information Criterion (AIC) [8] or the Bayesian Information Criterion (BIC) [9]. This has the distinct advantage of making it feasible in settings where the complete design matrix may have many thousands of regressors; such problems are becoming increasingly common in bioinformatics, e.g. microarray analysis and genome wide association studies. It remains to select  $\nu$ ; setting  $\nu = 1$  renders (10) the minimum variance unbiased estimator of  $\tau$ , and is therefore one sensible choice.

*Remark 1: Coding.* To construct the uniform code it has been assumed that the observed signal power  $\mathbf{y}'\mathbf{y}$  is known by the receiver. Alternatively, one can design a preamble code to transmit this quantity to the receiver, making the entire message decodable. As all regression models will require this preamble code it may safely be ignored during model selection for moderate sample sizes.

*Remark 2: Alternative Prior.* An attempt at forming 'uninformative' priors for linear regression models in the MML literature was made in [2]. Here, an additive uniform prior density over the coefficients was chosen to reflect the belief that the higher order coefficients will account for the remaining variance that is unexplained by the already selected lower order coefficients. However, such a prior is not uniform over the feasible set of the regression parameters and depends on an arbitrary ordering of the coefficients.

## 4 Linear Regression with the $g$ -Prior

Instead of the uniform prior used in the  $\text{MML}_{Lu}$  criterion, we now consider a multivariate Gaussian prior over the regression coefficients. Dropping  $\gamma$  for brevity, this results in the following hierarchy:

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \tau\mathbf{I}_n) \quad (12)$$

$$\boldsymbol{\beta} \sim N_p\left(\mathbf{0}_p, m(\mathbf{X}'\mathbf{X})^{-1}\right) \quad (13)$$

where both  $m > 0$  and  $\tau$  are hyperparameters that must be estimated from the data. The type of prior considered here is known as Zellner's  $g$ -prior [10]. Coding of the assertion now proceeds by first transmitting estimates for  $\boldsymbol{\alpha} = (m, \tau)$ , and then transmitting the regression parameters  $\boldsymbol{\beta}$  given the hyperparameters. This is further detailed in [7].

The negative log-likelihood of the data  $\mathbf{y}^n$  given the parameters  $(\boldsymbol{\beta}, \tau)$  is given by (3). The Fisher information for  $\boldsymbol{\beta}$  now requires a correction as the hyperparameter  $m$  is estimated from the data, and may be arbitrarily small leading to problems with the uncorrected MML87 approximation. Following the procedure described in [1] (pp. 236–237), the corrected Fisher information is formed by treating the prior  $\pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}|m)$  as a posterior of some uninformative prior  $\pi_0(\boldsymbol{\beta})$  and  $p$  prior data samples all set to zero, with design matrix  $\mathbf{X}_0$  satisfying  $\mathbf{X}'_0\mathbf{X}_0 = (\tau/m)(\mathbf{X}'\mathbf{X})$ . The corrected Fisher information is

$$|\mathbf{J}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{\alpha})| = \left(\frac{\tau + m}{\tau m}\right)^p |\mathbf{X}'\mathbf{X}| \quad (14)$$

Substituting (3), (13) and (14) into (1), and minimising the resultant codelength for  $\boldsymbol{\beta}$  yields the following MML87 estimates:

$$\hat{\boldsymbol{\beta}}_{87}(\mathbf{y}^n|\boldsymbol{\alpha}) = \left(\frac{m}{m + \tau}\right) (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \left(\frac{m}{m + \tau}\right) \hat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{y}^n) \quad (15)$$

Using the procedure described in Section 2, the profile message length, say  $I_{\hat{\boldsymbol{\beta}}}$ , evaluated at  $\hat{\boldsymbol{\beta}}_{87}(\mathbf{y}^n|\boldsymbol{\alpha})$  up to constants not depending on  $\boldsymbol{\alpha}$  is

$$\frac{n}{2} \log \tau + \frac{p}{2} \log \left(\frac{\tau + m}{\tau}\right) + \left(\frac{1}{2\tau}\right) \mathbf{y}'\mathbf{y} - \left(\frac{m}{2\tau(m + \tau)}\right) \xi(\mathbf{y}^n)$$

where  $\xi(\mathbf{y}^n) = \hat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{y}^n)'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{y}^n)$  is the fitted sum of squares. Noting that  $\text{E}[\mathbf{y}'\mathbf{y}] = n\tau + mp$  and  $\text{E}[\xi(\mathbf{y}^n)] = p(\tau + m)$ , the entries of the Fisher information matrix for the hyperparameters  $\boldsymbol{\alpha}$  are

$$\text{E} \left[ \frac{\partial^2 I_{\hat{\boldsymbol{\beta}}}}{\partial m^2} \right] = \text{E} \left[ \frac{\partial^2 I_{\hat{\boldsymbol{\beta}}}}{\partial m \partial \tau} \right] = \frac{p}{2(m + \tau)^2} \quad (16)$$

$$\text{E} \left[ \frac{\partial^2 I_{\hat{\boldsymbol{\beta}}}}{\partial \tau^2} \right] = \frac{p}{2(m + \tau)^2} + \frac{n - p}{2\tau^2} \quad (17)$$

yielding the Fisher information

$$|\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| = \frac{(n-p)p}{4\tau^2(m+\tau)^2} \quad (18)$$

The hyperparameters  $m$  and  $\tau$  are given the uninformative prior

$$\pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \propto \tau^{-\nu} \quad (19)$$

where  $\nu$  is specified *a priori*. Substituting (3), (13), (14), (18) and (19) into (2) and minimising the resultant codelength with respect to  $\boldsymbol{\alpha}$  yields

$$\begin{aligned} \hat{\tau}_{87}(\mathbf{y}^n) &= \frac{\mathbf{y}'\mathbf{y} - \xi(\mathbf{y}^n)}{n-p+2\nu-2} \\ \hat{m}_{87}(\mathbf{y}^n) &= \left( \frac{\xi(\mathbf{y}^n)}{\delta} - \hat{\tau}_{87}(\mathbf{y}^n) \right)_+ \end{aligned}$$

where  $\delta = \max(p-2, 1)$  and  $(\cdot)_+ = \max(\cdot, 0)$  as  $m$  may never be negative. When  $\hat{m}_{87}(\mathbf{y}^n) > 0$ , the complete minimum message length for the data, parameters and the hyperparameters is given by

$$\begin{aligned} I_g(\mathbf{y}^n|\gamma) &= \left( \frac{n-p+2\nu-2}{2} \right) (\log \hat{\tau}_{87}(\mathbf{y}^n) + 1) + \frac{p-2}{2} \log \frac{\xi(\mathbf{y}^n)}{\delta} \\ &\quad + \frac{1}{2} \log(n-p)p^2 + \text{const} \end{aligned} \quad (20)$$

where  $I_g(\mathbf{y}^n|\gamma) \equiv I_g(\mathbf{y}^n, \hat{\boldsymbol{\beta}}_{87}(\mathbf{y}^n|\hat{\boldsymbol{\alpha}}), \hat{\boldsymbol{\alpha}}_{87}(\mathbf{y}^n)|\gamma)$ . Alternatively, when  $\hat{m}_{87}(\mathbf{y}^n) = 0$  or  $p = 0$ , we instead create a ‘no effects’ design matrix  $\mathbf{X}$  comprising a single regressor such that  $\mathbf{X}'\mathbf{X} = 0$  which yields the codelength

$$I_g(\mathbf{y}^n|\emptyset) = \left( \frac{n+2\nu-4}{2} \right) (\log \hat{\tau}_{87}(\mathbf{y}^n) + 1) + \frac{1}{2} \log(n-1) + \frac{1}{2} + \text{const} \quad (21)$$

for the ‘null’ model  $\gamma = \emptyset$ . The codes (20)–(21) are referred to as  $\text{MML}_g$ ; as they depend on the design matrix only through  $\hat{\tau}_{87}(\mathbf{y}^n)$  and  $\xi(\mathbf{y}^n)$  they are invariant under all full rank affine transformations of  $\mathbf{X}$ . As in the case of  $\text{MML}_u$ , the  $\text{MML}_g$  criterion is of the same computational complexity as AIC and BIC, and is therefore also suitable for application to high dimensional problems.

*Remark 3: Minimality.* The  $\text{MML}_{87}$  estimators (15) are minimax with respect to squared error loss for all  $p > 2$ , assuming the choice of  $\nu = 2$  [11]. Furthermore, such estimators dominate the standard least squares estimators for all  $p > 2$ , and generally outperform them even for  $p < 3$ .

*Remark 4:* Given that the least squares estimates are always minimax, it may be preferable to use  $\text{MML}_u$  when  $p < 3$  and  $\text{MML}_g$  otherwise. This requires a calibration of the two codelengths involving a more efficient coding of the hyperparameter in  $\text{MML}_u$  and is a topic for future work.

## 5 Coding the Model Index

The previous criteria ignored the requirement for stating  $\gamma$ , which denotes the regressors included in the model. However, in order to decode the message, a receiver needs to know  $\gamma$ . In nested model selection, such as polynomial regression, a reasonable choice is to use a uniform prior over the maximum number of regressors,  $q > 0$ , under consideration with codelength

$$I(\gamma) = \log(q + 1) \quad (22)$$

If one is instead considering the all subsets regression setting, a prior that treats each combination of subsets of size  $p$  as equally likely may be chosen [12]. This prior yields the codelength

$$I(\gamma) = \log \binom{q}{p} + \log(q + 1) \quad (23)$$

An alternative prior is to treat *all* combinations of regressors as equally likely; this is equivalent to using a Bernoulli prior with probability of including a regressor set to  $(1/2)$ . However, it is not difficult to show that for moderate  $q$ , such a prior results in codelengths that are longer than those obtained by (23) for almost all combinations of regressors.

Once a suitable code for the model index is chosen, regressor selection is performed by solving

$$\hat{\gamma} = \arg \min_{\gamma} \{I(\mathbf{y}^n | \gamma) + I(\gamma)\}$$

where  $I(\mathbf{y}^n | \gamma)$  is the codelength of the regression model specified by  $\gamma$ ; for example, the  $\text{MML}_u$  (11) or the  $\text{MML}_g$  (20)–(21) codes.

## 6 Discussion and Results

The  $\text{MML}_u$  and  $\text{MML}_g$  criteria are now compared against two state of the art MDL linear regression criteria, denoted NML [4] and  $g\text{MDL}$  [13], and the KICc [14] method, on both synthetic and real data. The hyperparameter  $\nu$  was set to  $\nu = 1$  for the  $\text{MML}_u$  criterion, and to  $\nu = 2$  for the  $\text{MML}_g$  criterion. It has been previously shown that NML,  $g\text{MDL}$  and KICc regularly outperform the well known AIC and BIC (and allied criteria) and we therefore do not include them in the tests. The NML criterion  $I_{\text{NML}}(\mathbf{y}^n)$  for  $p > 0$ , up to constants, is given by:

$$\left(\frac{n-p}{2}\right) \log \frac{\mathbf{y}'\mathbf{y} - \xi(\mathbf{y}^n)}{n} + \frac{p}{2} \log \frac{\xi(\mathbf{y}^n)}{p} - \log \Gamma\left(\frac{n-p}{2}\right) - \log \Gamma\left(\frac{p}{2}\right) \quad (24)$$

The  $\text{MML}_u$  and  $\text{MML}_g$  criteria are clearly similar in form to (24); in fact, using a Jeffrey's prior over  $\boldsymbol{\alpha}$  yields a codelength that differs from (24) by  $(1/2) \log p + O(1)$ . This is interesting given that the NML criterion is derived with the aim of

Criterion		Sample Size							
		25	50	75	100	125	150	200	500
$\hat{p} < p$	MML <sub>u</sub>	15.67	1.64	0.09	0.00	0.00	0.00	0.00	0.00
	MML <sub>g</sub>	20.82	2.03	0.12	0.01	0.00	0.00	0.00	0.00
	NML	15.62	0.87	0.03	0.00	0.00	0.00	0.00	0.00
	gMDL	8.110	0.35	0.01	0.00	0.00	0.00	0.00	0.00
	KICc	32.64	1.50	0.04	0.00	0.00	0.00	0.00	0.00
$\hat{p} = p$	MML <sub>u</sub>	62.27	86.07	91.24	93.10	94.25	94.90	95.78	97.70
	MML <sub>g</sub>	<b>65.02</b>	88.83	<b>93.12</b>	<b>94.62</b>	<b>95.41</b>	<b>95.91</b>	<b>96.60</b>	<b>98.09</b>
	NML	63.38	84.68	89.27	91.26	92.53	93.38	94.53	96.97
	gMDL	64.48	82.17	87.19	89.45	91.15	92.15	93.51	96.46
	KICc	62.19	<b>89.33</b>	89.94	89.44	89.27	88.94	88.80	88.43
$\hat{p} > p$	MML <sub>u</sub>	22.07	12.30	8.674	6.896	5.755	5.100	4.219	2.301
	MML <sub>g</sub>	14.16	9.141	6.759	5.374	4.589	4.087	3.397	1.909
	NML	21.00	14.46	10.70	8.741	7.474	6.621	5.468	3.026
	gMDL	27.41	17.48	12.80	10.549	8.847	7.854	6.487	3.536
	KICc	5.170	9.162	10.08	10.56	10.73	11.06	11.20	11.57
Error	MML <sub>u</sub>	113.8	10.38	3.932	2.286	1.625	1.242	0.852	0.294
	MML <sub>g</sub>	50.86	7.144	<b>3.195</b>	<b>2.001</b>	<b>1.470</b>	<b>1.149</b>	<b>0.806</b>	<b>0.286</b>
	NML	95.38	12.34	4.914	2.871	2.026	1.472	0.957	0.309
	gMDL	136.1	15.69	5.955	3.345	2.302	1.637	1.035	0.319
	KICc	<b>18.37</b>	<b>5.607</b>	3.614	2.770	2.313	1.890	1.414	0.584

**Table 1.** Polynomial order selected by the criteria (expressed as percentages) and squared error in estimated coefficients

producing minimax regret codes rather than using a formal Bayesian argument. The MML<sub>u</sub> criterion may also be rendered even closer to NML by taking the tighter bound  $\xi(\mathbf{y}^n)$  when constructing the feasible parameter set  $A$ . The gMDL approach is derived from the Bayesian mixture code using the  $g$ -prior and is thus also closely related to MML<sub>g</sub>. The main differences lie in the coding of the hyperparameters, and the fact that the explicit two-part nature of MML<sub>g</sub> yields invariant point estimates for  $\beta$ .

## 6.1 Polynomial Regression

An example application of the newly developed MML criteria is to the problem of polynomial order selection. Following [14] and [4], the simple polynomial basis  $x^i$  for  $(i = 0, \dots, q)$  are used. Datasets of various sample sizes  $25 \leq n \leq 500$  were generated from the true model

$$y^* = x^3 - 0.5x^2 - 5x - 1.5$$

with the design points uniformly generated in  $[-3, 3]$ . The variance  $\tau$  was chosen to yield a signal-to-noise ratio of one. For every dataset, each criterion was asked to select a nested polynomial model up to maximum degree  $q = 10$ , and for each sample size the experiment was repeated  $10^5$  times; the model  $p = 0$  was not considered in this test. As the problem is one of nested model selection, the prior (22) was chosen to encode  $\gamma$ . The results are given in Table 1, where the error is the squared  $\ell_2$  norm of the difference between estimated and true coefficients. In terms of order selection, MML<sub>g</sub> is uniformly superior for all  $n$ , although for

Training Sample		Model Selection Criteria				
		MML <sub>u</sub>	MML <sub>g</sub>	NML	gMDL	KICc
Housing	25	71.509	<b>61.922</b>	69.602	74.842	66.111
	50	36.635	<b>36.340</b>	36.918	37.075	36.460
	100	29.383	29.624	29.332	29.135	<b>29.053</b>
	200	26.162	26.424	26.031	<b>25.907</b>	26.025
	400	24.299	24.304	24.315	24.330	<b>24.217</b>
Diabetes	25	4819.2	<b>4445.0</b>	4952.5	5136.6	4457.6
	50	3843.8	3851.2	3945.0	3822.6	<b>3684.0</b>
	100	3364.2	3385.3	3361.2	3339.3	<b>3293.8</b>
	200	3173.3	3199.6	3166.7	3154.2	<b>3085.7</b>
	400	3052.7	3052.8	3047.3	3045.2	<b>3031.8</b>
Concrete	25	227.41	<b>221.20</b>	225.80	225.01	237.02
	50	149.25	<b>147.46</b>	148.65	148.52	148.46
	100	123.65	<b>122.90</b>	123.82	123.92	123.04
	200	114.50	114.37	114.56	114.62	<b>114.33</b>
	400	111.64	<b>111.59</b>	111.67	111.67	111.62

**Table 2.** Squared prediction errors for three real datasets estimated by cross-validation

large  $n$  the performance of all the MML/MDL criteria is similar. For small  $n$ , the MML<sub>g</sub> criterion achieves the best error of all the coding based approaches, followed by MML<sub>u</sub>; however, both are slightly inferior to KICc for  $n = 25$  and  $n = 50$ . This is due to the conservative nature of KICc. As the sample size grows, KICc achieves significantly lower correct order selection scores – this is not surprising as KICc is not consistent. Interestingly, even though KICc is asymptotically efficient it still attains a larger error than MML<sub>g</sub> at large sample sizes. Of the two MDL criteria, the gMDL criterion appears more prone to overfitting than NML and subsequently attains poorer order selection and error scores for almost all sample sizes.

## 6.2 Real Datasets

Three real datasets (two from the UCI machine learning repository [15], and one previously analysed in [16] among others) were used to assess the performance of the new MML criteria. The datasets were: (1) the Boston housing data ( $q = 14$ ,  $n = 506$ ), (2) the diabetes data ( $q = 10$ ,  $n = 442$ ) and (3) the concrete compressive strength dataset ( $q = 9$ ,  $n = 1030$ ). Each dataset was randomly divided into a training and testing sample, and the five criteria were asked to choose a suitable subset (including the ‘no effects’ model) of the candidate regressors from the training sample. The testing sample was subsequently used to assess the predictive performance of the criteria, measured in terms of squared error. Each test was repeated  $10^3$  times. As this was an all-subsets regression problem, the prior (23) was used for all coding based methods, with  $q$  set appropriately depending on the dataset used. The results are presented in Table 2.

Both MML criteria perform well for small sample sizes ( $n \leq 25$ ) and tend to perform marginally worse than the MDL criteria for larger sample sizes. Of the two MML criteria, the MML<sub>g</sub> criterion appears slightly superior to MML<sub>u</sub>. The KICc criterion is competitive which is not unexpected given that it is an

efficient criterion and is expected to perform well for prediction. An interesting point to note is that when  $MML_g$  outperforms NML and  $gMDL$ , the difference in performance can be relatively large; in contrast, when the MML criteria obtain higher prediction errors than the MDL criteria, the difference in prediction is minimal. The  $MML_g$  criterion thus appears to offer better protection against overfitting when there is little signal available (i.e. small sample size or large noise) while trading off little performance as the signal strength increases.

## References

1. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. First edn. Information Science and Statistics. Springer (2005)
2. Fitzgibbon, L.J., Dowe, D.L., Allison, L.: Univariate polynomial inference by Monte Carlo message length approximation. In: 19th International Conference on Machine Learning (ICML'2002), Sydney, Australia (2002) 147–154
3. Viswanathan, M., Wallace, C.S.: A note on the comparison of polynomial selection methods. In: Uncertainty 99: The Seventh International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, Florida (1999) 169–177
4. Rissanen, J.: Information and Complexity in Statistical Modeling. First edn. Information Science and Statistics. Springer (2007)
5. Wallace, C.S., Boulton, D.M.: An information measure for classification. Computer Journal **11**(2) (August 1968) 185–194
6. Wallace, C.S., Freeman, P.R.: Estimation and inference by compact coding. Journal of the Royal Statistical Society (Series B) **49**(3) (1987) 240–252
7. Makalic, E., Schmidt, D.F.: Minimum message length shrinkage estimation. Statistics & Probability Letters **79**(9) (May 2009) 1155–1161
8. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6) (December 1974) 716–723
9. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2) (March 1978) 461–464
10. Zellner, A.: Applications of Bayesian analysis in econometrics. The Statistician **32**(1–2) (1983) 23–34
11. Sclove, S.L.: Improved estimators for coefficients in linear regression. Journal of the American Statistical Association **63**(322) (June 1968) 596–606
12. Roos, T., Myllymäki, P., Rissanen, J.: MDL denoising revisited. IEEE Transactions on Signal Processing **57**(9) (2009) 3347–3360
13. Hansen, M.H., Yu, B.: Model selection and the principle of minimum description length. Journal of the American Statistical Association **96**(454) (2001) 746–774
14. Seghouane, A.K., Bekara, M.: A small sample model selection criterion based on Kullback's symmetric divergence. IEEE Trans. Sig. Proc. **52**(12) (2004) 3314–3323
15. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
16. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. The Annals of Statistics **32**(2) (2004) 407–451