MML Logistic Regression with Translation and Rotation Invariant Priors

Enes Makalic and Daniel F. Schmidt

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

Abstract. Parameters in logistic regression models are commonly estimated by the method of maximum likelihood, while the model structure is selected with stepwise regression and a model selection criterion, such as AIC or BIC. There are two important disadvantages of this approach: (1) maximum likelihood estimates are biased and infinite when the data is linearly separable, and (2) the AIC and BIC model selection criteria are asymptotic in nature and tend to perform well only when the sample size is moderate to large. This paper introduces a novel criterion, based on the Minimum Message Length (MML) principle, for parameter estimation and model selection of logistic regression models. The new criterion is shown to outperform maximum likelihood in terms of parameter estimation, and outperform both AIC and BIC in terms of model selection using both real and artificial data.

Consider the logistic regression model for explaining data $\mathbf{y} \in \mathbb{R}^n$ given a matrix of covariates $\mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')'$

$$p(\mathbf{y}|\mathbf{X}, \alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{1}{1 + \exp(-y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta}))} \right) \tag{1}$$

where the target variable $y_i \in \{-1, +1\}$, $\mathbf{x}_i \in \mathbb{R}^p$ $(i = 1, 2, \ldots, n)$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of logistic regression coefficients and $\alpha \in \mathbb{R}$ is the intercept parameter. The p regression coefficients $\boldsymbol{\beta}$ and the intercept parameter α determine the probability of the target variable $y_i = \pm 1$. The task in logistic regression is to estimate the $(p + 1)$ parameters $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})' \in \mathbb{R}^{p+1}$ and select a subset of the p predictor variables that is associated with the target.

Logistic regression is the most commonly used model in epidemiology and social science for analysis of data with a binary outcome. The parameter coefficients are often estimated by the method of maximum likelihood

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{y}) = (\hat{\alpha}(\mathbf{y}), \hat{\boldsymbol{\beta}}'(\mathbf{y}))' = \arg\max_{\alpha, \beta} \left\{ \prod_{i=1}^{n} \frac{1}{1 + \exp(-y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta}))} \right\} \tag{2}$$

which effectively sets the free parameters to the values that maximise the likelihood or, equivalently, the log-likelihood of the observed data. However, the

M. Thielscher and D. Zhang (Eds.): AI 2012, LNCS 7691, pp. 878–889, 2012.

maximum likelihood estimates for logistic regression are biased and can be infinite in small samples and when the data is linearly separable. To select which predictors are associated with the target, one combines maximum likelihood with best subset selection or forward/backward regression and applies a model selection criterion, such as Akaike's Information Criterion (AIC) [1] or the Bayesian Information Criterion (BIC) [2].

This paper considers minimum message length (MML) [3] parameter estimation and model selection in logistic regression analysis. We derive a new Bayesian model selection criterion that is defined even when the data is linearly separable, requires no user specified parameters and exhibits good performance in small samples. A procedure for obtaining MML parameter estimates is introduced and the new estimates are shown to exhibit better prediction performance than the traditional maximum likelihood estimates. The new model selection criterion is then empirically evaluated against two popular criteria, AIC and BIC, and demonstrates excellent performance. This is a remarkable result. By minimising the MML codelength function we obtain parameter estimates that are superior to maximum likelihood and bias–corrected maximum likelihood estimates, especially in small samples with correlated covariates. Further, minimising the same codelength function yields a model selection criteria that outperforms both AIC and BIC.

1 Minimum Message Length (MML)

Minimum message length (MML) [4,5,6,3] principle of inductive inference states that the best model is one which results in the best compression, or shortest codelength, of the data. The compressed codelength comprises two parts: (1) the *assertion*, stating a model for the data from a set of candidate models, and (2) the *detail* which encodes the data using the model that was named in the assertion. The assertion and the detail are commonly denoted as $I_{87}(\boldsymbol{\theta})$ and $I_{87}(\mathbf{y}|\boldsymbol{\theta})$ respectively. The most commonly used form of MML is the Wallace–Freeman approximation [6], or the MML87 approximation, which states that the codelength, $I_{87}(\mathbf{y}, \boldsymbol{\theta})$, of model $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ and data $\mathbf{y} \in \mathbb{R}^n$ is

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}) = \underbrace{-\log \pi(\boldsymbol{\theta}) + \frac{1}{2}\log |\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| + \frac{k}{2}\log \kappa_k}_{I_{87}(\boldsymbol{\theta})} + \underbrace{\frac{k}{2} - \log p(\mathbf{y}|\boldsymbol{\theta})}_{I_{87}(\mathbf{y}|\boldsymbol{\theta})} \tag{3}$$

where $\pi(\cdot)$ denotes a prior distribution over the support $\boldsymbol{\Theta}$, $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function, $\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the $(k \times k)$ Fisher information matrix and $\kappa_k > 0$ is a dimensionality constant. Following Wallace (p. 237, [3]), the dimensionality constant is well approximated by

$$\frac{k}{2}(\log \kappa_k + 1) \approx -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(k\pi) + \psi(1) \tag{4}$$

where $\psi(\cdot)$ is the digamma function. MML is a Bayesian principle and requires stating a prior density over the free model parameters. Under MML87, the model

$\hat{\boldsymbol{\theta}}_{87}(\mathbf{y})$ which minimises (5) is chosen as the most a posteriori likely explanation of the data $\mathbf{y}$, in view of the chosen prior density $\pi(\cdot)$.

The Wallace–Freeman approximation is derived under the following assumptions: (1) the log-likelihood is approximately quadratic at the maximum, (2) the Fisher information matrix is positive definite over the support $\boldsymbol{\Theta}$ and (3) the prior density is locally continuous and 'slowly' varying around the region determined by $|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})|$. Assumptions (2) and (3) do not hold for logistic regression models with the priors selected in Section 2.1. This is because the chosen prior density has a singularity at the origin and the Fisher information is semi-positive definite; that is, the Fisher information tends to zero as the regression parameters tend to infinity. The combination of these two factors may result in a breakdown of the codelength approximation and hence nonsensical codelengths.

To alleviate these issues, we use the "small sample" codelength approximation suggested by Wallace ([3], pp. 235–236)

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}) = \underbrace{\frac{1}{2} \log \left(1 + \frac{|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \mathbf{I}_k|\kappa_k^k}{\pi(\boldsymbol{\theta})^2} \right)}_{I_{87}(\boldsymbol{\theta})} + \underbrace{\frac{k}{2} - \log p(\mathbf{y}|\boldsymbol{\theta})}_{I_{87}(\mathbf{y}|\boldsymbol{\theta})} \tag{5}$$

Further, we add an identity matrix to the Fisher information matrix to prevent the parameters from shooting off to infinity and prevent the codelength approximation from breaking down. This change to the Fisher information matrix makes little to no difference around the 'true' MML estimate, but significantly improves the codelength accuracy when the parameters are large (that is, the Fisher information is nearly singular). The new codelength is unfortunately no longer invariant under model transformation due to the modification to the Fisher information matrix. However, this violation appears minor since the models that actually minimise the codelength are virtually unaffected by the change to the Fisher information.

If the number of parameters is fixed and does not grow with the sample size, the MML87 estimates asymptotically converge to maximum likelihood as the sample size $n \to \infty$. Furthermore, the MML87 codelength is asymptotically equivalent to the Bayesian Information Criterion (BIC) as $n \to \infty$. In contrast to maximum likelihood, MML87 estimates are statistically consistent in certain difficult inference problems; for example, the Neyman–Scott problem [7] and the factor analysis model. The MML87 estimator also tends to improve upon maximum likelihood in problems where the data size is small to moderate. Thus, the MML criterion may be viewed as a small–sample version of BIC that also features improved parameter estimates.

2 MML Logistic Regression

Standard Bayesian inference of logistic regression models requires specifying a prior distribution over the parameters and sampling from the corresponding posterior distribution given the observed data. This has traditionally been done

using Markov Chain Monte Carlo (MCMC) techniques with Gaussian approximations, some variant of the Metropolis–Hastings sampler [8] or numerical integration techniques. More recently, researchers have examined MCMC techniques with alternative representations of the logistic function and Gaussian scale-mixture priors [9]. A common representation of the logistic function is to model the data $\mathbf{y} \in \mathbb{R}^n$ as a thresholded version of some underlying continuous random variable. Examples include the the random-utility construction of the logistic model by Holmes and Held [10], the Z–distribution framework of Gramacy and Polson [11] and the Polya–gamma representation of Polson et al. [12].

In the Bayesian framework, the prior distribution hierarchy commonly follows the local/global shrinkage framework described by Polson and Scott [13]. The prior distribution over the regression coefficients is taken to be the Gaussian distribution with a prior variance hyperparameter, which is given a separate hyperprior. The functional form of the hyperprior allows the generation of several widely used distributions for the regression coefficients through Gaussian-scale mixtures; examples of distributions that may be generated in this setting include the Student t-distribution or the double exponential distribution used in Bayesian LASSO regression [14], among others.

Although the MML principle is Bayesian by design, the MML approximation (5) that is used in this paper does not require any sampling from the posterior distribution of the parameters. Instead, one must specify: (1) the negative log-likelihood function, (2) the determinant of the Fisher information matrix, and (3) a prior distribution over the free model parameters for the logistic regression model. The fully-specified model $\boldsymbol{\theta}(\mathbf{y})$ which minimises the codelength approximation given by (5) is then the best MML model in light of the chosen priors. The negative log-likelihood function and the Fisher information for logistic regression models are

$$-\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta}))), \tag{6}$$

$$|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| = |(\mathbf{1}_n, \mathbf{X})'\mathbf{V}(\boldsymbol{\theta})(\mathbf{1}_n, \mathbf{X})|\,, \tag{7}$$

where $\mathbf{1}_n = (1, 1, \ldots, 1)'$ is an n-dimensional vector of ones, $\mathbf{V}(\boldsymbol{\theta}) = \text{diag}(\mu_1(1-\mu_1), \mu_2(1-\mu_2), \ldots, \mu_n(1-\mu_n))$ is an $(n \times n)$ diagonal matrix, $\mu_i = 1/(1 + \exp(-\alpha - \mathbf{x}_i'\boldsymbol{\beta}))$ and $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})' \in \mathbb{R}^k$ denotes the $k = (p + 1)$ vector of free parameters. It remains to specify the prior distribution over the regression coefficients.

2.1 Prior Distribution

Almost all Bayesian approaches to logistic regression use a Gaussian prior distribution (or Gaussian-scale mixtures) over regression coefficients $\boldsymbol{\theta} \in \mathbb{R}^k$. This is done primarily out of mathematical convenience. However, the Gaussian distribution is not invariant under rotations and translations of the decision hyperplane implied by the logistic model. We argue that this is not desirable and

seek a prior that is invariant under all translations and rotations of the decision hyperplane.

The decision boundary in logistic regression models is

$$\alpha + \mathbf{x}'\boldsymbol{\beta} = \tilde{\alpha}(1 + \mathbf{x}'\tilde{\boldsymbol{\beta}}) = 0, \tag{8}$$

which is a $(p-1)$-dimensional hyperplane embedded inside the p-dimensional data space induced by the predictor matrix $\mathbf{X}$, assuming $\mathbf{X}$ is full rank. Note that the logistic regression model is now parameterised in terms of $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\boldsymbol{\beta}}')' \in \mathbb{R}^k$, where the parameter transformation function $F : \mathbb{R}^k \mapsto \mathbb{R}^k$ maps $\boldsymbol{\theta} \in \mathbb{R}^k$ to $\tilde{\boldsymbol{\theta}} = F(\boldsymbol{\theta}) = (F_0(\boldsymbol{\theta}), F_1(\boldsymbol{\theta}), \ldots, F_p(\boldsymbol{\theta}))'$ and

$$\tilde{\alpha} = F_0(\boldsymbol{\theta}) = \alpha \tag{9}$$

$$\tilde{\beta}_j = F_j(\boldsymbol{\theta}) = \beta_j/\alpha \quad (j = 1, 2, \ldots, p), \tag{10}$$

provided $\alpha \neq 0$. It is clear from (8) that $\tilde{\alpha} \in \mathbb{R}$ does not affect the location and orientation of the decision boundary which is solely determined by the regression coefficients $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p$.

A prior distribution over $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p$ that does not favor any orientation or position of the decision boundary is

$$\pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}}) = \frac{\Gamma(p/2) r_0}{2\pi^{p/2}} (\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}})^{-(p+1)/2}, \quad (||\tilde{\boldsymbol{\beta}}||_2 \geq r_0 > 0), \tag{11}$$

where $\Gamma(\cdot)$ is the gamma function and r_0 is a lower limit on the ℓ_2-norm of the parameter vector $\tilde{\boldsymbol{\beta}}$ (discussed below). The prior distribution was originally derived in the context of feed-forward multilayer perceptron networks [15]. Let $\mathbf{I}_p$ denote a $(p \times p)$ identity matrix. It is straightforward to show that $-\log \pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}})$ is a strictly concave function of $\tilde{\boldsymbol{\beta}}$ since the $(p \times p)$ Hessian matrix

$$-\frac{\partial^2 \log \pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}'} = \left(\frac{p+1}{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}\right) \left(\mathbf{I}_p - \left(\frac{2}{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}\right) \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}'\right) \tag{12}$$

has a strictly negative determinant

$$\left| -\frac{\partial^2 \log \pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}'} \right| = -\left(\frac{p+1}{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}\right)^p < 0, \quad \tilde{\boldsymbol{\beta}} \in \mathbb{R}^p. \tag{13}$$

The constant r_0 may be set to the radius of the minimum volume enclosing hyperball of the predictor matrix $\mathbf{X}$; this can be efficiently computed for moderate p using the fast algorithm in [16]. Alternatively, one may set $r_0^2 = 1/(\mathbf{t}'\mathbf{t})$ where $\mathbf{t} \in \mathbb{R}^p$ is the point closest to the origin through which the decision hyperplane passes. This procedure was recommended by Toussaint et al. ([15], Section 3.1). Section 2.2 discusses how to set the constant r_0 in the current paper.

It remains to specify a prior distribution over the parameter $\tilde{\alpha} \in \mathbb{R}$. We opt for the scale invariant prior

$$\pi_{\tilde{\alpha}}(\tilde{\alpha}) = \frac{a}{2\tilde{\alpha}^2}, \quad \tilde{\alpha} \in [a, \infty). \tag{14}$$

where $a > 0$. Note that Toussaint et al. ([15], Section 3.1) advocate a maximum entropy prior for $\tilde{\alpha}$ in neural networks using the maximal slope of the decision boundary as testable information. The maximum entropy prior however introduces an extra hyperparameter which is not easily eliminated. Since $\tilde{\alpha}$ is a common parameter to all models, the choice of prior is not expected to significantly alter the results and conclusions of this paper.

To summarise, the complete prior distribution over the k free parameters $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\boldsymbol{\beta}}')'$ is

$$\pi_{\tilde{\boldsymbol{\theta}}}(\tilde{\boldsymbol{\theta}}) = \pi_{\tilde{\alpha}}(\tilde{\alpha})\pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}}). \tag{15}$$

where the prior distributions for $\tilde{\alpha}$ and $\tilde{\boldsymbol{\beta}}$ are assumed to be conditionally independent and are given by (14) and (11), respectively.

2.2 MML Logistic Regression Criterion

It remains to specify the complete MML codelength (5) for the logistic regression model as a function of the new parameters $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\boldsymbol{\beta}}')' \in \mathbb{R}^k$ (see Section 2.1). The negative log-likelihood and the Fisher information are now given by

$$-\log p(\mathbf{y}|\mathbf{X}, \tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\tilde{\alpha}(1 + \mathbf{x}'\tilde{\boldsymbol{\beta}})))) \tag{16}$$

$$|\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\tilde{\boldsymbol{\theta}})| = |\mathbf{J}_{\mathbf{T}}{}'\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})\mathbf{J}_{\mathbf{T}}| = \tilde{\alpha}^{2p}|(\mathbf{1}_n, \mathbf{X})'\mathbf{V}(\tilde{\boldsymbol{\theta}})(\mathbf{1}_n, \mathbf{X})| \tag{17}$$

where $\mathbf{J}_{\mathbf{T}}$ is the $(k \times k)$ Jacobian transformation matrix

$$\mathbf{J}_{\mathbf{T}} = \begin{pmatrix} 1 & \mathbf{0}_p' \\ \tilde{\boldsymbol{\beta}} & \tilde{\alpha}\mathbf{I}_p \end{pmatrix}, \quad |\mathbf{J}_{\mathbf{T}}| = \tilde{\alpha}^p \tag{18}$$

and $\mathbf{V}(\tilde{\boldsymbol{\theta}}) = \mathbf{V}(\boldsymbol{\theta})$ (see Section 2). The complete MML codelength for logistic regression is

$$I_{87}(\mathbf{y}, \tilde{\boldsymbol{\theta}}) = \frac{1}{2}\log\left(1 + \frac{|\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\tilde{\boldsymbol{\theta}}) + \mathbf{I}_k|\kappa_k^k}{\pi(\tilde{\boldsymbol{\theta}})^2}\right) + \sum_{i=1}^{n} \log(1 + \exp(-y_i(\tilde{\alpha}(1 + \mathbf{x}'\tilde{\boldsymbol{\beta}})))) + \frac{k}{2}$$

where the prior density and the Fisher information are given in (15) and (17) respectively. The constant κ_k^k is approximated using (4). We set the prior parameters a and r_0 to the parameter estimates that minimise the codelength. That is, we choose $a = \hat{\tilde{\alpha}}(\mathbf{y})$ and $r_0 = \hat{\tilde{\boldsymbol{\beta}}}(\mathbf{y})'\hat{\tilde{\boldsymbol{\beta}}}(\mathbf{y})$.

We have thus far ignored the important requirement for stating which of the p predictors are used in the model under consideration. In this paper, we choose a prior distribution that treats each possible subset of regressors of size q $(0 < q \leq p)$ as equally likely. The new prior adds a term of size

$$\log(p + 1) + \log\binom{p}{q}$$

to the codelength (19). The first term states the number $q(\leq p)$ of predictors which are in the model, while the second term names which of the p predictors are selected. Recall that a model with q predictors always includes an additional intercept term, represented by the parameter $\tilde{\alpha} \in \mathbb{R}$, and thus has $(q+1)$ free parameters.

2.3 Parameter Estimation

It is well known that maximum likelihood parameter estimates for logistic regression (2) are biased away from the point $\boldsymbol{\theta} = \mathbf{0}_k$ and are infinite if the data is linearly separable [17]. In order to remove or lessen the bias, some amount of parameter shrinkage towards the origin is necessary. The bias arises due to the curvature and unbiasedness of the score function (that is, the derivative of the log-likelihood). A common method of reducing the bias is that of Firth [18] where, instead of maximising the log-likelihood, one maximises the penalized log-likelihood

$$\hat{\boldsymbol{\theta}}_{\mathrm{FR}}(\mathbf{y}) = (\hat{\alpha}(\mathbf{y}), \hat{\boldsymbol{\beta}}'(\mathbf{y}))' = \arg\max_{\alpha,\boldsymbol{\beta}} \left\{ \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) + \frac{1}{2} \log |\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| \right\} \tag{19}$$

where the penalty is equal to the determinant of the Fisher information matrix. This penalized log-likelihood introduces bias into the score function by an amount that depends on the curvature of the log-likelihood. The end result for exponential families is the removal of $O(n^{-1})$ bias for the canonical parameter.

Firth's penalized log-likelihood estimates are not model space invariant, since bias itself is not invariant. Shen and Gao [19] extend the approach and add a second, ridge regression, penalty to the log-likelihood function to improve maximum likelihood estimates in case of highly correlated predictors. A potential disadvantage of both penalized likelihood estimates is that commonly used model selection criteria, such as AIC and BIC, are only derived assuming maximum likelihood estimates and may need modification for penalized regression procedures. In contrast, the MML approach allows parameter estimation and model selection within the same framework. The authors recommend [20] for an overview of other similar approaches to bias reduction in logistic regression models.

The MML parameter estimates, $\hat{\tilde{\boldsymbol{\theta}}}_{87}(\mathbf{y})$ can be obtained by minimising the total codelength (19) with respect to the parameters $\tilde{\boldsymbol{\theta}}$, that is

$$\hat{\tilde{\boldsymbol{\theta}}}_{87}(\mathbf{y}) = \arg\min_{\tilde{\alpha},\tilde{\boldsymbol{\beta}}} \left\{ I_{87}(\mathbf{y}, \tilde{\boldsymbol{\theta}}) \right\}. \tag{20}$$

The Fisher information term $\log |\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\tilde{\boldsymbol{\theta}})|$ and the prior distribution $\pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}})$ are strictly concave functions and unbounded below. The negative log-likelihood function is strictly convex and bounded below. In this paper, the parameter estimates are obtained using the MATLAB `fminunc` optimisation function. Preliminary simulation experiments initially showed that minimising the codelength

with respect to both $\tilde{\alpha}$ and $\tilde{\boldsymbol{\beta}}$ can sometimes result in numerical issues, especially as $\tilde{\alpha} \to 0$ and the parameter transformation becomes singular. A possible remedy is to fix $\tilde{\alpha}$ to equal, say, $\hat{\alpha}_{\mathrm{FR}}(\mathbf{y})$, and estimate the p regression coefficients by minimising the codelength. This removes all issues with numerical optimisation and yields valid codelengths provided the initial estimate of $\tilde{\alpha}$ is reasonable. The effect of the prior distribution $\pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}})$ on the MML codelength is to shrink the maximum likelihood parameters towards the origin of the coordinate system. Given the tendency of maximum likelihood to overestimate the regression coefficients, we expect the MML estimates to exhibit less bias (see Section 3.1).

It is interesting to note that some penalized regression procedures can be interpreted within the MML framework. The bias correction suggested by Firth, for example, amounts to maximising the posterior distribution of the parameters assuming a Jeffreys' prior distribution. This is equivalent to using a constant Fisher information when computing MML87 estimates.

3 Results and Discussion

3.1 Parameter Estimation

This section compares the prediction performance of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{y})$, the MML parameter estimator $\hat{\tilde{\boldsymbol{\theta}}}_{87}(\mathbf{y})$ and Firth's penalized maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathrm{FR}}(\mathbf{y})$ in logistic regression models. The three estimators are defined in (2), (20) and (19) respectively. In the spirit of reproducible research, all MATLAB simulation code will be made available on the authors' web pages[1]. Recall that

$$\mu_i = \frac{1}{1 + \exp(-y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta}))} \quad (i = 1, 2, \ldots, n), \tag{21}$$

denotes the probability of datum $y_i = \pm 1$, and let $\hat{\mu}_i$ denote an estimate of μ_i inferred by one of the three aforementioned estimators. Similarly, let $\hat{y}_i$ denote the predicted class of the datum y_i. The estimates will be compared on the area under the receiver operating characteristic curve (AUC) and

$$\mathrm{CA} = \frac{1}{n}\sum_{i=1}^{n} \delta(y_i, \hat{y}_i) \tag{22}$$

$$\mathrm{KL} = -\frac{1}{n}\sum_{i=1}^{n} \left(\mu_i \log\left(\frac{\hat{\mu}_i}{\mu_i}\right) + (1 - \mu_i)\log\left(\frac{1 - \hat{\mu}_i}{1 - \mu_i}\right)\right), \tag{23}$$

where CA denotes classification accuracy, KL denotes Kullback–Leibler divergence [21] and $\delta(y_i, \hat{y}_i) = 1$ if and only if $(y_i = \hat{y}_i)$, otherwise $\delta(y_i, \hat{y}_i) = 0$. These three metrics will measure the prediction error of the estimators under consideration.

[1] `www.emakalic.org/blog` and `www.dschmidt.org`

Table 1. Parameter estimation performance measured by median classification accuracy (CA), area under the ROC curve (AUC) and KL divergence (KL) computed for maximum likelihood $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{y})$, penalized maximum likelihood $\hat{\boldsymbol{\theta}}_{\mathrm{FR}}(\mathbf{y})$ and minimum message length $\hat{\boldsymbol{\theta}}_{87}(\mathbf{y})$ estimators

n	ρ	$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{y})$			$\hat{\boldsymbol{\theta}}_{\mathrm{FR}}(\mathbf{y})$			$\hat{\boldsymbol{\theta}}_{87}(\mathbf{y})$		
		CA	AUC	KL	CA	AUC	KL	CA	AUC	KL
	0.0	72.16	75.01	8.72	72.45	81.11	0.58	72.43	83.10	0.57
	0.2	77.79	81.85	6.59	77.77	87.25	0.47	85.45	94.71	0.40
25	0.5	80.11	84.72	5.66	79.92	89.37	0.43	90.44	97.74	0.30
	0.7	81.00	85.51	5.39	80.46	90.10	0.41	91.92	98.45	0.26
	0.9	81.53	86.21	5.17	81.01	90.57	0.41	93.12	98.85	0.24
	0.0	78.11	85.54	0.87	78.48	87.60	0.49	77.50	87.69	0.52
	0.2	83.36	86.93	4.89	84.06	92.95	0.36	86.89	95.30	0.31
50	0.5	85.97	89.36	4.13	86.42	94.88	0.30	91.15	97.83	0.24
	0.7	86.99	90.39	3.78	87.24	95.44	0.29	92.63	98.48	0.21
	0.9	87.68	91.16	3.52	87.84	95.83	0.27	93.66	98.85	0.19
	0.0	81.47	90.30	0.44	81.47	90.31	0.41	81.50	90.39	0.40
	0.2	87.17	95.11	0.35	87.26	95.23	0.30	87.95	95.69	0.28
100	0.5	89.67	96.40	0.41	89.89	96.97	0.24	91.26	97.72	0.21
	0.7	90.35	95.16	1.33	90.69	97.45	0.22	92.69	98.38	0.18
	0.9	90.97	94.08	2.43	91.33	97.79	0.20	93.84	98.85	0.15
	0.0	82.94	91.49	0.38	82.93	91.49	0.37	82.95	91.50	0.37
	0.2	88.67	96.07	0.27	88.67	96.07	0.26	88.84	96.17	0.26
250	0.5	91.39	97.70	0.21	91.39	97.71	0.20	91.76	97.88	0.19
	0.7	92.30	98.16	0.20	92.33	98.17	0.18	92.85	98.40	0.17
	0.9	92.94	98.46	0.19	92.95	98.47	0.17	93.89	98.82	0.15

The test procedure for comparing the estimators is now described. For each test, a training data set comprising the predictor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ was generated from a multivariate Gaussian distribution $\mathbf{X} \sim \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma})$, where the entries of the variance–covariance matrix are $\boldsymbol{\Sigma}_{i,i} = 1$ for all $(i = 1, 2, \ldots, p)$ and $\boldsymbol{\Sigma}_{i,j} = \rho$ whenever $(i \neq j)$; that is, $\boldsymbol{\Sigma}$ is a variance–covariance matrix with ones on the diagonal and ρ everywhere else. This is expected to produce significant correlation in the covariates as $|\rho| \to 1$. The training data $\mathbf{y} \in \mathbb{R}^n$ was generated using the predictor matrix $\mathbf{X}$ and $(\alpha, \boldsymbol{\beta}) = (0', \mathbf{1}')'$ for the unknown parameters $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')' \in \mathbb{R}^{p+1}$. The maximum likelihood, penalized maximum likelihood and the MML estimators were then used to estimate the unknown parameters $\boldsymbol{\theta}$ given the training data. For the the MML estimate, $\hat{\tilde{\alpha}}(\mathbf{y})$ is set to the Firth estimate (see Section 2.3). The performance of the three estimators was compared using the three metrics CA, AUC and KL computed from a new test data set with sample size $(m = 10^5)$. In all tests, $(p = 10)$ predictor variables were used to generate the predictor matrices. The entire procedure was repeated for 2000 iterations for the following values of (n, ρ): $n \in \{25, 50, 100, 250\}$ and $\rho \in \{0.0, 0.2, 0.5, 0.7, 0.9\}$. The results are shown in Table 1.

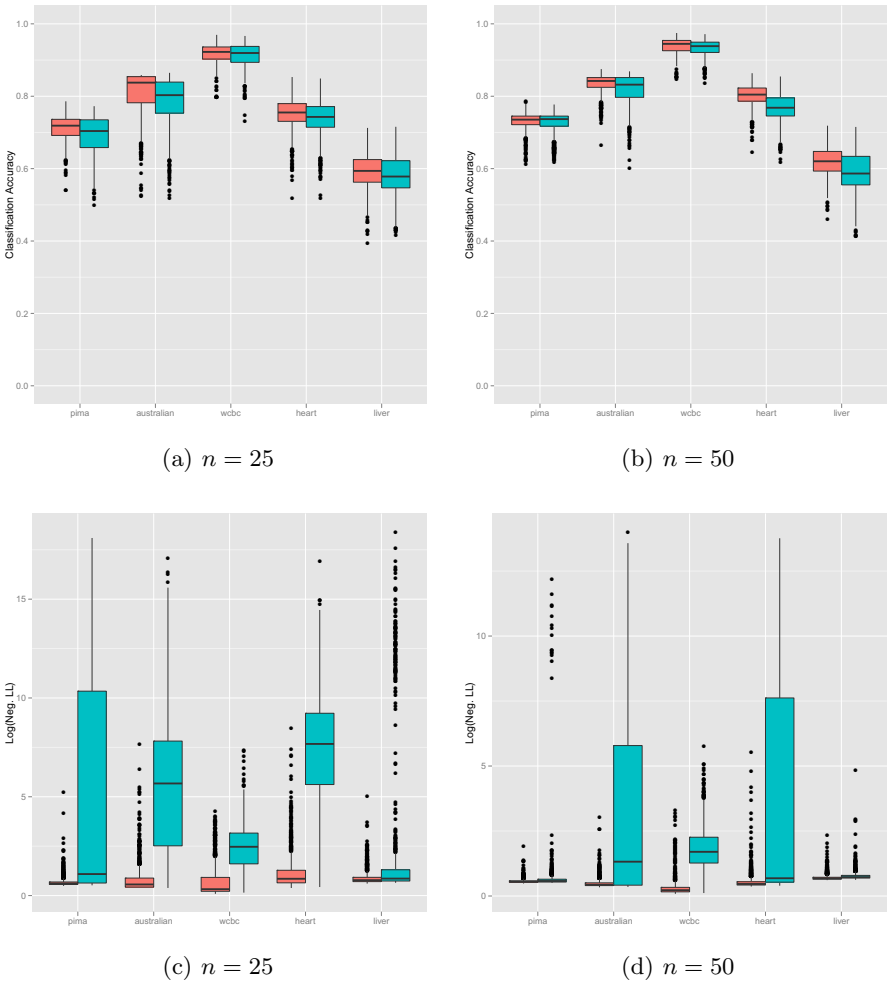
It is clear that both the MML and Firth's penalized likelihood estimator substantially improve the prediction performance of maximum likelihood, especially when the sample size is small ($n < 100$). This is not surprising as such data is often linearly separable and maximum likelihood tends to estimate $\boldsymbol{\beta} \to \pm\infty$. The MML estimator and Firth's penalized likelihood estimator exhibit similar prediction performance for larger sample sizes ($n \geq 100$) and for all values of correlation ρ. However, the MML estimator has significantly better prediction error as measured by all three metrics for $n = \{25, 50\}$, especially as $\rho \to 1$. This indicates that the MML estimator is able to better estimate the underlying 'degrees of freedom' of the model when the data is highly correlated. For larger sample sizes, all three estimators performed equally well under the metrics considered.

3.2 Model Selection

The MML logistic regression criterion is now compared against AIC and BIC in terms of model selection performance on five real data sets. The data sets were: `pima` (8 predictors, 768 samples), `australian` (15 predictors, 690 samples), `wcbc` (10 predictors, 683 samples), `liver` (6 predictors, 345 samples) and `heart` (13 predictors, 270 samples). The `pima`, `wcbc` and `liver` data sets were obtained from the UCI Machine Learning Repository, while the remaining data sets were downloaded from StatLOG.

For each data set, random samples of $n = 25$ and $n = 50$ data were used for training, while the remaining data was used for estimating prediction performance. Prior to each test iteration, the predictor variables were normalized to have zero mean and length $\mathbf{x}_j'\mathbf{x} = n$ for all $(j = 1, 2, \ldots, p)$. Initially, all-subset selection was used to generate candidate models. However, the computational complexity of examining all 2^p models renders any kind of exhaustive empirical comparison extremely difficult. Consequently, the elastic net procedure [22] (function `lassoglm` in MATLAB with parameter $\alpha = 0.95$) was used to generate candidate models – the elastic net parameter estimates were ignored. Two metrics were computed from the test data: (a) classification accuracy, and (b) negative log-likelihood. The complete test procedure was repeated for 10^3 iterations for each data set. The results are shown in Figure 1. Note, the AIC score resulted in inferior performance in comparison to BIC in virtually all tests considered and was hence omitted.

In terms of median classification accuracy, both MML and BIC perform similarly across all data sets and sample sizes considered. However, the models selected by the MML criterion have significantly smaller variance in terms of classification accuracy than those selected by BIC. This is clearly visible for $n = 50$ training samples where the variance in classification accuracy of the MML models is approximately half of the BIC models. In terms of negative log-likelihood, the MML criterion has resulted in clearly superior models to BIC for both $n = 25$ and $n = 50$ training samples.



(a) $n = 25$ (b) $n = 50$

(c) $n = 25$ (d) $n = 50$

Fig. 1. Classification accuracy and negative log-likelihood computed from test data for MML (left boxplot) and BIC (right boxplot) for sample sizes $n = 25$ and $n = 50$

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (1974)
2. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6(2), 461–464 (1978)
3. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length, 1st edn. Information Science and Statistics. Springer (2005)
4. Wallace, C.S., Boulton, D.M.: An information measure for classification. Computer Journal 11(2), 185–194 (1968)

5. Wallace, C., Boulton, D.: An invariant Bayes method for point estimation. Classification Society Bulletin 3(3), 11–34 (1975)
6. Wallace, C.S., Freeman, P.R.: Estimation and inference by compact coding. Journal of the Royal Statistical Society (Series B) 49(3), 240–252 (1987)
7. Dowe, D.L., Wallace, C.S.: Resolving the Neyman-Scott problem by minimum message length. In: Proc. 28th Symposium on the Interface, Sydney, Australia. Computing Science and Statistics, vol. 28, pp. 614–618 (1997)
8. Metropolis, A.W., Rosenbluth, M.N., Rosenbluth, A.H., Teller, E.: Equations of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087–1092 (1953)
9. Andrews, D.F., Mallows, C.L.: Scale mixtures of normal distributions. Journal of the Royal Statistical Society (Series B) 36(1), 99–102 (1974)
10. Holmes, C.C., Held, L.: Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Analsyis 1(1), 145–168 (2006)
11. Gramacy, R.B., Polson, N.G.: Simulation-based regularized logistic regression. Bayesian Analsyis 7(3) (to appear, 2012)
12. Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using polya-gamma latent variables. arXiv:1205.0310
13. Polson, N.G., Scott, J.G.: Shrink globally, act locally: Sparse Bayesian regularization and prediction. In: Bayesian Statistics, vol. 9 (2010)
14. Park, T., Casella, G.: The Bayesian lasso. Journal of the American Statistical Association 103(482), 681–686 (2008)
15. van Toussaint, U., Gori, S., Dose, V.: Invariance priors for Bayesian feed-forward neural networks. Neural Networks 19(10), 1550–1557 (2006)
16. Gärtner, B.: Fast and Robust Smallest Enclosing Balls. In: Nešetřil, J. (ed.) ESA 1999. LNCS, vol. 1643, pp. 325–338. Springer, Heidelberg (1999)
17. Albert, A., Anderson, J.A.: On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1), 1–10 (1984)
18. Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika 80(1), 27–38 (1993)
19. Shen, J., Gao, S.: A solution to separation and multicollinearity in multiple logistic regression. J. Data Sci. 6(4), 515–531 (2008)
20. Bull, S.B., Mak, C., Greenwood, C.M.T.: A modified score function estimator for multinomial logistic regression in small samples. Computational Statistics & Data Analysis 39(1), 57–74 (2002)
21. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics 22(1), 79–86 (1951)
22. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society (Series B) 67(2), 301–320 (2005)