Minimum Message Length Ridge Regression for Generalized Linear Models

Daniel F. Schmidt and Enes Makalic

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

Abstract. This paper introduces an information theoretic model selection and ridge parameter estimation criterion for generalized linear models based on the minimum message length principle. The criterion is highly general in nature, and handles a range of target distributions, including the normal, binomial, Poisson, geometric and gamma distributions. Estimation of the regression parameters, the ridge hyperparameter and the set of covariates associated with targets is all performed within the same framework by minimisation of the message length. Experiments on simulated and real data suggest that the criterion is competetive with, and often superior to, the corrected Akaike information criterion in terms of both parameter estimation and model selection tasks.

1 Introduction

In conventional Gaussian-linear regression modelling we make the assumption that the targets $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ are normally distributed, with variance τ, and mean μ_i given by

$$\mu_i = \bar{\mathbf{x}}_i \boldsymbol{\beta} + \alpha, \tag{1}$$

where $\bar{\mathbf{x}}_i' \in \mathbb{R}^k$ is a vector of features, $\boldsymbol{\beta} \in \mathbb{R}^k$ is a vector of regression coefficients, and $\alpha \in \mathbb{R}$ is the intercept parameter. It is typically the case that we do not believe the targets to be normally distributed; for example, the targets may be non-negative integers or binary variables. The generalized linear model (GLM) [1] framework was developed to easily extend linear models to alternative target distributions. In this paper we restict attention to distributions which satisfy

$$\mathrm{E}\left[y|\mu\right] = \mu, \tag{2}$$

$$\mathrm{var}[y|\mu, \phi] = \phi\, v(\mu), \tag{3}$$

where $\phi > 0$ is a dispersion parameter which in many cases will simply be equal to one, and $v(\cdot)$ is a variance function dependent only on the mean μ. Defining $\boldsymbol{\psi} = (\alpha, \boldsymbol{\beta}')'$ as the vector of regression coefficients and $\eta_i(\boldsymbol{\psi}) \equiv \eta_i = \bar{\mathbf{x}}_i \boldsymbol{\beta} + \alpha$ as the linear predictor, the GLM approach specifies $f(\mu_i) = \eta_i$, where $f(\cdot)$ is called a *link function*; that is, a GLM specifies the conditional mean as a suitable

(monotic, isomorphic) function of the linear predictor. The function $f^{-1}(\eta_i) = \mu_i$ is usually known as the inverse-link function.

In general, the regression coefficients α and $\boldsymbol{\beta}$ are unknown, and we only have access to the data $\mathbf{y}$ and the covariates $\mathbf{X} = (\bar{\mathbf{x}}_1', \ldots, \bar{\mathbf{x}}_n')'$. The task is then to estimate the regression coefficients on the basis of the data alone. There exists a large range of estimation strategies available for GLMs, and a particularly popular approach is *ridge regression* [2]. This is a regularisation procedure that is known to improve estimation accuracy in the presence of colinearity in the covariates. The (generalized) ridge regression procedure estimates $\alpha, \boldsymbol{\beta}$ by solving

$$\left\{\hat{\alpha}, \hat{\boldsymbol{\beta}}\right\} = \underset{\alpha \in \mathbb{R}, \boldsymbol{\beta} \in S(c)}{\arg\min} \left\{ -\sum_{i=1}^{n} \log p(y_i | \mu_i; \phi) \right\} \tag{4}$$

where $p(\cdot)$ is the chosen target distribution and $S(c)$ is the set of permissible regression coefficients, defined by

$$S(c) = \left\{ \boldsymbol{\beta} \in \mathbb{R}^k : \boldsymbol{\beta}' \boldsymbol{\Sigma} \boldsymbol{\beta} \leq c \right\}, \tag{5}$$

with $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ a positive-definite matrix. The hyperparameter c determines the amount of "freedom" the estimator has to fit the data; for a sufficiently large choice of c the ridge estimator reduces to the regular maximum likelihood estimator, while smaller values of c result in estimates that are "shrunk" towards the origin. It is usual to estimate c by minimisation of an information criteria such as Akaike's information criterion (AIC), or by a resampling procedure such as cross-validation. It is possible to interpret the ridge estimator in a Bayesian manner, in which the regularisation term arises due to the choice of a multivariate normal prior distribution over the regression coefficients $\boldsymbol{\beta}$.

In this paper we exploit this Bayesian interpretation to use the minimum message length (MML) principle to estimate the regularisation hyper-parameter; furthermore, because of the nature of the MML principle, we can also use the same criterion to perform feature selection, i.e., to choose which columns of $\mathbf{X}$ are associated with the targets $\mathbf{y}$. The result is a *single, highly general criterion for the statistical inference of generalized linear models that is applicable to wide range of target distributions*, and has excellent performance in terms of parameter estimation and model selection.

2 Inference by Minimum Message Length

Minimum message length (MML) [3–5] is an information theoretic principle of inductive inference based on the connections between statistical inference and data compression. The key idea underlying the MML principle is that if a statistical model compresses data, then the model has (with high probability) captured regularities and structure in the data. The MML principle advocates selecting the model that most compresses the data (i.e., the one with the shortest "message length") as the most plausible explanation of the data. As any compressed

representation of data must also be decompressable, the details of the statistical model used to encode the data must also be part of the compressed data string. Thus, more complex models inflate the message length by a greater amount, and this acts to naturally balance model complexity against the goodness of fit of the model, and automatically guards against the problem of overfiting the data.

In general, the calculation of the exact (strict) message length is an NP-hard problem [6]. There exists a range of approximations to the exact message length that less computationally intensive [5]; the most widely used of these is the Wallace–Freeman approximation (MML87) [4]. Let $\boldsymbol{\theta} \in \Theta$ denote the continuous parameters of a statistical model, $p(\mathbf{y}|\boldsymbol{\theta})$ denote the likelihood of the data $\mathbf{y}$ conditional on the parameters $\boldsymbol{\theta}$, and let $\pi(\boldsymbol{\theta})$ denote a Bayesian prior distribution over Θ that will be used to model the continuous parameters. The MML87 message length for data $\mathbf{y}$ and model $\boldsymbol{\theta}$ is given by

$$I(\mathbf{y}, \boldsymbol{\theta}) = -\log p(\mathbf{y}|\boldsymbol{\theta}) + \frac{1}{2}\log|\mathbf{J}(\boldsymbol{\theta})| - \log \pi(\boldsymbol{\theta}) + c(k) \tag{6}$$

where $\mathbf{J}(\boldsymbol{\theta})$ is the Fisher information matrix, k is the number of continuous model parameters and

$$c(k) = -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(k\pi) - 0.5772.$$

To estimate a model using MML87, we search for the $\boldsymbol{\theta}$ that minimises (6). Under certain regularity conditions of the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ and prior distribution $\pi(\boldsymbol{\theta})$ the MML87 message length is very close to the exact strict message length [5]. The aim of this paper is to apply the MML87 approximation to the problem of ridge estimation and model selection in the context of generalized linear models.

Ridge estimation in the MML framework is equivalent to allowing the prior distribution to depend on a *hyperparameter*, and extending the estimation procedure to include this new hyperparameter. Previous work [7] has shown that inference of hyperparameters may be done within the MML87 framework, and this technique has been applied to linear regression with a normal target distribution and a special choice of ridge prior in [8]. MML has been previously applied to linear models with normal targets [5] and binomial targets [9], and both these cases essentially depend on special types of ridge priors. To some extent, the MML criterion presented in this paper generalises this previous work, as it allows for general ridge estimation and a large number of target distributions.

3 MML GLM Ridge Regression

To compute message lengths using the MML87 approximation (6) we require: (i) the negative log-likelihood function; (ii) prior distributions over all parameters; and (iii) an appropriate Fisher information matrix. Define the full vector of parameters for a GLM as $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', \phi)'$, where ϕ may be constrained to $\phi = 1$ for some target distributions, and define $\boldsymbol{\psi} = (\alpha, \boldsymbol{\beta}')'$ as the vector of regression parameters. It is usual to assume that the targets are independent

random variables, conditional on the features, so that the likelihood function can be factorised into the product

$$p(\mathbf{y}|\boldsymbol{\theta};\mathbf{X}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta};\bar{\mathbf{x}}_i). \tag{7}$$

To implement ridge regression within a Bayesian context the required prior distribution for the $\boldsymbol{\beta}$ coefficients is a multivariate normal with mean $\mathbf{0}_k$ and variance-covariance matrix $(\phi/\lambda)\boldsymbol{\Sigma}^{-1}$. Scaling the covariance matrix by the dispersion parameter ϕ greatly simplifies the resulting estimates of α and $\boldsymbol{\beta}$ as they become independent of the estimate of ϕ. As the origin holds no special meaning for the intercept we choose a uniform distribution for α. The priors for α and $\boldsymbol{\beta}$, conditional on ϕ and λ are:

$$\pi(\boldsymbol{\psi}|\phi,\lambda) = \pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\phi,\lambda) \cdot \pi_{\alpha}(\alpha|\phi), \tag{8}$$

$$\pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\phi,\lambda) = \left(\frac{\lambda}{2\pi\phi}\right)^{\frac{k}{2}} \cdot |\boldsymbol{\Sigma}|^{\frac{1}{2}} \cdot \exp\left(-\frac{\lambda\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{2\phi}\right), \tag{9}$$

$$\pi_{\alpha}(\alpha|\phi) \propto \frac{1}{\sqrt{\phi}}. \tag{10}$$

Due to the fact we condition on ϕ in (9) and (10), we may first estimate α, $\boldsymbol{\beta}$, and then subsequently estimate ϕ (if required). The prior for α is improper, and must technically be restricted to some subset of $\mathbb{R}$; the particular choice of subset is unimportant as the α parameter is common to all GLMs and the normalisation term will simply increase all message lengths by a constant amount. Priors for ϕ are discussed in Section 3.1.

In the ridge regression framework, the regularisation hyper-parameter λ is not considered known *a priori*; rather, it is estimated from the data along with the other model parameters. This can introduce some problems into the standard MML87 message length, as the assumption of a "flat" prior distribution is violated when λ becomes very large, and the resulting normal distribution becomes tightly concentrated around the origin. To address this problem we use the "corrected" form of the Fisher information matrix that takes into account the curvature of the prior. To correct the Fisher information matrix, Wallace proposed a clever procedure in the case of conjugate likelihood and prior distributions, in which the model parameters are treated as "fake" data, and the Fisher information is calculated using both the real and "fake" data (see [5], pp. 236–237 for further details).

The likelihood (7) is not conjugate with the prior distribution (9). However, it is well known that the likelihood of many common GLMs can be approximated around some point, $\boldsymbol{\psi}_0 = (\alpha_0, \boldsymbol{\beta}_0')'$, by a multivariate normal distribution with appropriate mean and covariance matrix; such approximations form the basis of the efficient iteratively reweighted least squares procedure for maximum likelihood estimation of GLM regression coefficients. Define $\boldsymbol{\mu}_0 = f^{-1}(\mathbf{X}\boldsymbol{\beta}_0 + \alpha_0\mathbf{1}_n)$; the approximate negative log-posterior for $\boldsymbol{\psi}$, up to constants independent of $\boldsymbol{\psi}$,

is then given by

$$-\log p(\boldsymbol{\psi}|\mathbf{y},\phi,\lambda) \approx \left(\frac{1}{2\phi}\right)(\mathbf{z}(\boldsymbol{\mu}_0)-\mathbf{X}\boldsymbol{\beta}-\alpha\mathbf{1}_n)'\mathbf{W}(\boldsymbol{\mu}_0)(\mathbf{z}(\boldsymbol{\mu}_0)-\mathbf{X}\boldsymbol{\beta}-\alpha\mathbf{1}_n) + \left(\frac{\lambda}{2\phi}\right)\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}, \tag{11}$$

where $\mathbf{z}(\cdot)$ is a vector-valued function with entries

$$z_i(\boldsymbol{\mu}) = f^{-1}(\mu_i) + (y_i-\mu_i)\left(\frac{\partial f(\mu_i)}{\partial \mu_i}\right), \tag{12}$$

and $\mathbf{W}(\boldsymbol{\mu}_0) = \mathrm{diag}(\mathbf{w}(\boldsymbol{\mu}_0))$ is an $(n\times n)$ diagonal matrix, where $\mathbf{w}(\cdot)$ is a vector valued function with entries

$$W_{i,i}(\boldsymbol{\mu}) = \left(\frac{1}{v(\mu_i)}\right)\left(\frac{\partial f(\mu_i)}{\partial \mu_i}\right)^{-2}. \tag{13}$$

The functions $f(\mu_i)$, $f^{-1}(\eta_i)$ and $\partial f(\mu_i)/\partial\mu_i$ for several common choices of link function are given in Table 1, and the variance function $v(\mu_i)$ for a range of distributions is given in Table 2.

The likelihood term in the approximation (11) is conjugate with the normal prior density for the coefficients, and we may now view the prior $\pi_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\phi,\lambda)$ as the posterior of some uninformative prior $\pi_0(\boldsymbol{\beta})$ and a likelihood of k "prior samples", all equal to zero, with design matrix $\mathbf{X}_0$ satisfying $\mathbf{X}_0'\mathbf{X}_0 = (\lambda/\phi)\boldsymbol{\Sigma}$. This yields a "corrected" Fisher information matrix for the regression parameters $\boldsymbol{\psi}$ of the form

$$\mathbf{J}(\boldsymbol{\psi}|\phi,\lambda) = \left(\frac{1}{\phi}\right)\left((\mathbf{1}_n,\mathbf{X})'\,\mathbf{W}(\boldsymbol{\mu})\,(\mathbf{1}_n,\mathbf{X}) + \lambda\mathbf{S}\right), \tag{14}$$

where

$$\mathbf{S} = \begin{pmatrix} 0 & \mathbf{0}_k' \\ \mathbf{0}_k & \boldsymbol{\Sigma} \end{pmatrix}, \tag{15}$$

and $\boldsymbol{\mu} = f^{-1}(\mathbf{X}\boldsymbol{\beta}+\alpha\mathbf{1}_n)$. The "correction" has the effect of increasing the determinant of (14) for increasing λ; that is, the tighter the prior becomes around the origin, the larger the determinant of the corrected Fisher. In contrast, in the limit as $\lambda \to 0$ (and the normal prior (9) converges to a uniform distribution over $\boldsymbol{\beta}$) the corrected Fisher information reduces to the standard, "uncorrected" Fisher information.

3.1 Coding ϕ

Some target distributions require the coding of an extra dispersion parameter ϕ. This can be largely treated in a unified manner irrespective of the specific details of the target distribution by choosing the prior distribution $\pi_\phi(\cdot)$ to be the co-ordinate wise reference prior, i.e.,

$$\pi_\phi(\phi) \propto \sqrt{J(\phi)/n}, \tag{16}$$

	Link Function, $f(\mu_j)$	$\partial f(\mu_j)/\partial \mu_j$	Inverse Link, $f^{-1}(\eta_j)$
Identity	$\eta_j = \mu_j$	$\frac{\partial \eta_j}{\partial \mu_j} = 1$	$\mu_j = \eta_j$
Logit	$\eta_j = \log\left(\frac{\mu_j}{1-\mu_j}\right)$	$\frac{\partial \eta_j}{\partial \mu_j} = \frac{1}{\mu_j(1-\mu_j)}$	$\mu_j = \frac{1}{1+\exp(-\eta_j)}$
Log	$\eta_j = \log(\mu_j)$	$\frac{\partial \eta_j}{\partial \mu_j} = \frac{1}{\mu_j}$	$\mu_j = \exp(\eta_j)$

Table 1. Commonly used link functions and their derivatives and inverses; $\eta_j = \bar{\mathbf{x}}_j\boldsymbol{\beta}+\alpha$ is the linear predictor.

where $J(\phi)$ is the Fisher information for ϕ. Due to the fact that the distributions considered in Table 2 are parameterised in terms of orthogonal mean and dispersion parameters, the Fisher information for (ψ_i, ϕ) is zero for all $i = 1, \ldots, k+1$. The determinant of the full Fisher information matrix can be then written as the product

$$|\mathbf{J}(\boldsymbol{\theta}; \lambda)| = |\mathbf{J}(\boldsymbol{\psi}|\phi, \lambda)| \cdot J(\phi). \tag{17}$$

This decomposition, coupled with the choice of reference prior (16) dramatically simplifies the MML87 codelength for ϕ by cancelling the $J(\phi)$ terms present in the determinant of the Fisher information (17) and the prior (16).

3.2 Complete Message Length for a GLM

Two message length formula are required: one for the case in which $k > 0$ (i.e., there is at least one covariate included in the model), and one for the special case in which $k = 0$. We now cover these two cases seperately.

Message Length when $k > 0$. In this case, the model parameters are α, $\boldsymbol{\beta}$ and ϕ (if required). The prior (9) for $\boldsymbol{\beta}$ depends on λ, which is treated as an unknown hyperparameter that must be estimated from the data. Therefore, we also need to transmit λ to the receiver. There exists a procedure to determine the optimum codelength for hyperparameters in the MML framework [7], but it is difficult to apply to generalized linear models; instead, the codelength for λ is approximated by the usual asymptotic formula, i.e., $I(\lambda) = (1/2)\log n$. As there is only a single hyperparameter the suboptimal coding of λ is not expected to have any large effect on the resulting MML inferences.

The total number of free parameters is equal to $m = k + 2$ if ϕ is a free parameter, and $m = k + 1$ if ϕ is constrained to a constant for the target distribution under consideration (see Table 2 for details). Using (9), (10), (14),

	Link	PDF, $p(y_j\|\boldsymbol{\theta};\bar{\mathbf{x}}_j)$	$v(\mu_j)$	ϕ
Normal	Identity	$\left(\frac{1}{2\pi\tau}\right)^{\frac{1}{2}}\exp\left(-\frac{(y_j-\mu_j)^2}{2\tau}\right)$	1	τ
Binomial	Logit	$\mu_j^{y_j}(1-\mu_j)^{(1-y_j)}$	$\mu_j(1-\mu_j)$	1
Poisson	Log	$\frac{\mu_j^{y_j}\exp(-\mu_j)}{\Gamma(y_j+1)}$	μ_j	1
Geometric	Log	$\frac{\mu_j^{y_j}}{(\mu_j+1)^{y_j+1}}$	$\mu_j^2+\mu$	1
Gamma	Log	$\frac{y_j^{\frac{1}{\kappa}-1}\exp\left(-\frac{y_j}{\kappa\mu_j}\right)}{(\kappa\mu_j)^{\frac{1}{\kappa}}\,\Gamma\left(\frac{1}{\kappa}\right)}$	μ_j^2	κ
Inverse-Gaussian	Log	$\left(\frac{1}{2\pi\xi y_j^3}\right)^{\frac{1}{2}}\exp\left(-\frac{(y_j-\mu_j)^2}{2\xi\mu_j^2 y_j}\right)$	μ_j^3	ξ

Table 2. Commonly used distributions and their variance functions; μ_j is the appropriate mean function; we define $y_j \in \{0,1\}$ and $0^0 = 1$ for the binomial likelihood; κ is the inverse of the shape parameter in the case of the Gamma distribution and ξ is the inverse of the shape parameter in the case of the inverse-Gaussian distributions.

(16) and (17) in (6) yields

$$I(\mathbf{y},\boldsymbol{\theta},\lambda;\mathbf{X}) = -\log p(\mathbf{y}|\boldsymbol{\theta};\mathbf{X}) + \frac{1}{2}\log|\left((\mathbf{1}_n,\mathbf{X})'\,\mathbf{W}(\boldsymbol{\mu})\,(\mathbf{1}_n,\mathbf{X}) + \lambda\mathbf{S}\right)| \\ -\frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{k}{2}\log\left(\frac{2\pi}{\lambda}\right) + \left(\frac{\lambda}{2\phi}\right)\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} + (1/2)\log n + c(m) + \text{const} \tag{18}$$

where $\boldsymbol{\mu} = f^{-1}(\mathbf{X}\boldsymbol{\beta}+\alpha\mathbf{1}_n)$, $\mathbf{S}$ is given by (15), $\mathbf{W}(\boldsymbol{\mu})$ is given by (13) and const denotes constant terms independent of $\boldsymbol{\theta}$, λ and $\mathbf{y}$.

Message Length when $k = 0$. In this case, no covariates are being used to model the data $\mathbf{y}$ and the model parameters are simply the intercept α and the dispersion parameter ϕ (if required). The total number of parameters is then $m = 2$ if ϕ is a free parameter, or $m = 1$ otherwise. As $\boldsymbol{\beta}$ is not being transmitted, there is no requirement to transmit the hyperparameter λ, and the message length simplifies to

$$I(\mathbf{y},\alpha) = -\log p(\mathbf{y}|\boldsymbol{\theta};\mathbf{X}) + \frac{1}{2}\log\mathbf{1}_n'\mathbf{W}(\boldsymbol{\mu})\mathbf{1}_n + c(m) + \text{const}, \tag{19}$$

where $\boldsymbol{\mu} = f^{-1}(\alpha\mathbf{1}_n)$.

4 Estimating $\boldsymbol{\psi}$, $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$

Finding the exact estimates for $\hat{\boldsymbol{\psi}}$ that minimise (18) is computationally expensive due to the presence of the log-determinant of the Fisher information. To avoid this problem, the posterior mode, or maximum a posteriori (MAP) estimates will be used as a surrogate for the exact MML estimates; for moderate to large sample sizes, the difference between the MML and MAP estimates is expected to be small. Furthermore, for case of normal and gamma target distributions, the MAP and MML estimators exactly coincide. This is easy to verify by noting that the corrected Fisher information matrix (14) in both of these cases is independent of $\boldsymbol{\psi}$.

The posterior mode estimates may be obtained by using the well-known iteratively reweighted least-squares (IRLS) algorithm [10]. Although this algorithm is usually used to obtain the maximum likelihood estimates, it is easily adapted to find ridge estimates through the use of data augmentation. This is done by defining a new, augmented, design matrix

$$\mathbf{X}_A = \begin{pmatrix} \mathbf{1}_n & \mathbf{X} \\ \mathbf{0}_k & \operatorname{diag}(\sqrt{\mathbf{v}})\mathbf{E}' \end{pmatrix}, \tag{20}$$

where $\mathbf{v}$ are the eigenvalues of $\boldsymbol{\Sigma}$, and $\mathbf{E}$ is a matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}$. In the common case of $\boldsymbol{\Sigma} = \mathbf{I}_k$, we have $\mathbf{v} = \mathbf{1}_k$ and $\mathbf{E} = \mathbf{I}_k$.

The algorithm begins by initialising the estimate of the conditional mean vector with suitable starting values:

$$\hat{\boldsymbol{\mu}}_\lambda \leftarrow \begin{cases} \mathbf{y}/2 + 1/4 & \text{(Binomial)} \\ \mathbf{y} + 1/4 & \text{(Poisson)} \\ \mathbf{y} & \text{(Otherwise)} \end{cases} \tag{21}$$

The IRLS ridge algorithm then procedes as follows:

1. Form the augmented weight matrix and "data" vector using (12) and (13) :

$$\mathbf{W}_A(\hat{\boldsymbol{\mu}}_\lambda) \leftarrow \begin{pmatrix} \mathbf{W}(\hat{\boldsymbol{\mu}}_\lambda) & \mathbf{0}_{n\times k} \\ \mathbf{0}_{k\times n} & \sqrt{\lambda}\,\mathbf{I}_k \end{pmatrix}, \quad \mathbf{z}_A(\hat{\boldsymbol{\mu}}_\lambda) \leftarrow \begin{pmatrix} \mathbf{z}(\hat{\boldsymbol{\mu}}_\lambda) \\ \mathbf{0}_k. \end{pmatrix} \tag{22}$$

2. Update the estimates of the regression coefficients:

$$\hat{\boldsymbol{\psi}}_\lambda \leftarrow \left(\mathbf{X}_A'\mathbf{W}_A(\hat{\boldsymbol{\mu}}_\lambda)\mathbf{X}_A\right)^{-1}\mathbf{X}_A'\mathbf{W}_A(\hat{\boldsymbol{\mu}}_\lambda)\mathbf{z}_A(\hat{\boldsymbol{\mu}}_\lambda) \tag{23}$$

3. Update the estimate of the conditional mean vector:

$$\hat{\boldsymbol{\mu}}_\lambda \leftarrow f^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda + \hat{\alpha}_\lambda \mathbf{1}_n)$$

4. If the change in estimates is sufficiently small, terminate. Otherwise, go to Step 1.

An advantage of conditioning the prior (9) for $\boldsymbol{\beta}$ on ϕ is that the estimating equation for $\boldsymbol{\psi}$, given by (23), is independent of ϕ. As the choice of ϕ has no effect on the MAP estimate of the coefficients $\boldsymbol{\psi}$, we may first estimate $\boldsymbol{\psi}$ using the above procedure, and once a suitable estimate has been obtained, we may subsequently use it to estimate ϕ.

4.1 Estimating $\boldsymbol{\phi}$

Once we have obtained the MAP estimate for the model coefficients, $\hat{\boldsymbol{\psi}}_\lambda$, we may estimate the dispersion parameter ϕ, if necessary. An initial estimate for ϕ can then be obtained by minimising the approximate negative log posterior (11) for ϕ, yielding

$$\hat{\phi}_\lambda \approx \left(\frac{1}{n}\right)\left[(\mathbf{z}(\hat{\boldsymbol{\mu}}_\lambda) - \hat{\boldsymbol{\beta}}_\lambda\mathbf{X} - \hat{\alpha}_\lambda\mathbf{1}_n)'\mathbf{W}(\hat{\boldsymbol{\mu}}_\lambda)(\mathbf{z}(\hat{\boldsymbol{\mu}}_\lambda) - \hat{\boldsymbol{\beta}}_\lambda\mathbf{X} - \hat{\alpha}_\lambda\mathbf{1}_n) + \lambda\hat{\boldsymbol{\beta}}_\lambda'\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}_\lambda\right]. \tag{24}$$

In the case of the normal distribution ($\phi \equiv \tau$), this estimate is the exact MML estimate for the noise variance. In the case of the gamma and inverse Gaussian distributions, this estimate may be close for large sample sizes, but will not in general be equal to the exact MML estimate. We now detail how to find the MML estimate in these two cases.

Gamma regression In this case, $\phi \equiv \kappa$, which plays the role of the inverse of the shape parameter found in the usual parameterisation of the gamma distribution. Let $(\hat{\mu}_\lambda)_i$ denote the i-th co-ordinate of the conditional mean vector estimate $\hat{\boldsymbol{\mu}}_\lambda$. The MML estimate of κ may be obtained by minimising

$$\left(\frac{1}{\kappa}\right)\left[\sum_{i=1}^{n}\left(\frac{y_i}{(\hat{\mu}_\lambda)_i} + \log(\hat{\mu}_\lambda)_i\right) - \sum_{i=1}^{n}\log y_i\right] + \left(\frac{n}{\kappa}\right)\log\kappa + n\log\Gamma\left(\frac{1}{\kappa}\right). \tag{25}$$

Closed form solutions for the MML estimate do not exist, and they must be found numerically. The approximate estimate (24) is a suitable a starting point for a numerical minisation procedure.

Inverse Gaussian regression In this case, $\phi \equiv \xi$, which plays the role of the inverse of the shape parameter found in the usual parameterisation of the inverse Gaussian distribution. An exact estimate may be obtained by minimising the message length; this is given by

$$\hat{\xi}_\lambda = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{(y_i - (\hat{\mu}_\lambda)_i)^2}{y_i\,(\hat{\mu}_\lambda)_i^2}, \tag{26}$$

where $(\hat{\mu}_\lambda)_i$ denotes the i-th component of the conditional mean estimate $\hat{\boldsymbol{\mu}}_\lambda$.

4.2 Estimating $\boldsymbol{\lambda}$

The regularisation parameter λ may also be estimated from the data by minimisation of the message length. Due to the use of the "corrected" Fisher information matrix, the MML87 message length does not break down even for very large λ, and the MML estimate may be obtained by solving

$$\hat{\lambda} = \underset{\lambda\in\mathbb{R}_+}{\arg\min}\left\{I(\mathbf{y}, \hat{\boldsymbol{\theta}}_\lambda, \lambda; \mathbf{X})\right\},$$

where $\hat{\boldsymbol{\theta}}_\lambda = (\hat{\boldsymbol{\psi}}_\lambda', \hat{\phi}_\lambda)'$, and $\hat{\boldsymbol{\psi}}_\lambda$ and $\hat{\phi}_\lambda$ are the estimates for $\boldsymbol{\psi}$ and ϕ, conditional on λ, obtained using the procedures described Sections 4 and 4.1.

5 Selecting Covariates

One of the strengths of MML is that minimisation of the message length can be used to both estimate continuous model parameters, as well as perform model selection. In the setting of GLMs, the most common model selection problem is identifying which covariates from a design matrix are associated with the target. The MML ridge scheme developed in this paper can easily be adapted to perform model selection. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_q)$ denote the complete, $(n \times q)$ design matrix, where $\mathbf{x}_i \in \mathbb{R}^n$, let $\gamma \subset \{1, \ldots, q\}$ index a particular subset of covariates, and let $k_\gamma = |\gamma|$ denote the number of covariates in the subset. We can then define a sub-design matrix by

$$\mathbf{X}_\gamma = \left(\mathbf{x}_{\gamma_1}, \ldots, \mathbf{x}_{\gamma_{k_\gamma}}\right).$$

For the message to be decodable, the particular subset γ being used must encoded; a prior over Γ is therefore required. If nothing is known *a priori* about the likelihood of any covariate being included in the final model, a prior that treats all subset sizes equally likely is appropriate [8]. This yields a codelength of

$$I(\gamma) = \log \binom{q}{k_\gamma} + \log(q+1).$$

The MML estimate of γ is then found by solving

$$\hat{\gamma} = \underset{\gamma \in \Gamma}{\arg\min} \left\{ I(\mathbf{y}, \hat{\boldsymbol{\theta}}_\lambda, \hat{\lambda}; \mathbf{X}_\gamma) + I(\gamma) \right\},$$

where $I(\mathbf{y}, \hat{\boldsymbol{\theta}}_\lambda, \hat{\lambda}; \mathbf{X}_\gamma)$ is given by either (18) or (19), depending on k_γ.

6 Experiments

The MML GLM ridge criterion was compared to the corrected Akaike Information Criterion (AIC_c) in both parameter estimation and model selection experiments. The AIC_c has previously been shown to perform well when applied to regression models, even in the case of small sample sizes [11]. Given a particular λ, the AIC_c score for the model is

$$\text{AIC}_c(\mathbf{y}; \lambda, \mathbf{X}) = -\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\lambda; \mathbf{X}) + \hat{k}_\lambda \left(\frac{n}{n - \hat{k}_\lambda - 1} \right), \tag{27}$$

where

$$\hat{k}_\lambda = \text{Tr}\left(\mathbf{X} \left(\mathbf{X}_A' \mathbf{W}_A(\hat{\boldsymbol{\mu}}_\lambda) \mathbf{X}_A \right)^{-1} \mathbf{X}' \mathbf{W}(\hat{\boldsymbol{\mu}}_\lambda) \right)$$

is the degrees-of-freedom of the fitted regression model, $\mathbf{X}_A$ is the augmented design matrix given by (20), $\mathbf{W}(\hat{\boldsymbol{\mu}}_\lambda)$ is the weight matrix given by (13), $\mathbf{W}_A(\hat{\boldsymbol{\mu}}_\lambda)$ is given by (22), $\hat{\boldsymbol{\theta}}_\lambda$ are the MAP estimates of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\mu}}_\lambda = f^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda + \hat{\alpha}_\lambda \mathbf{1}_n)$. The AIC_c estimate of λ is found by minimising (27).

		$\rho = 0.1$		$\rho = 0.5$		$\rho = 0.9$	
	n	AIC_c	MML	AIC_c	MML	AIC_c	MML
Normal	25	0.66	0.47	0.57	0.45	0.31	0.35
	50	0.91	0.70	0.86	0.71	0.64	0.62
	100	0.96	0.85	0.95	0.84	0.83	0.79
	250	0.98	0.94	0.98	0.94	0.93	0.90
Binomial	25	0.03	0.05	0.02	0.06	0.02	0.03
	50	0.56	0.22	0.51	0.19	0.27	0.19
	100	0.82	0.69	0.77	0.64	0.63	0.56
	250	0.96	0.91	0.94	0.89	0.89	0.88
Poisson	25	0.77	0.58	0.83	0.62	0.46	0.39
	50	0.94	0.96	0.96	0.95	0.92	0.89
	100	1.00	1.00	0.99	0.98	0.97	0.97
	250	1.00	1.00	1.00	1.00	1.00	1.00

Table 3. Median ratios of the Kullback–Leibler (KL) divergences obtained by the MML and AIC_c estimates over the KL divergence obtained by the maximum likelihood estimates.

6.1 Parameter Estimation Simulations

The performance of both the MML ridge estimates and the AIC_c ridge estimates were compared to the maximum likelihood estimates on simulated data. At each of the $1,000$ iterations of the simulation, a vector of $k = 10$ "true" regression coefficients was sampled from a normal distribution, $\beta_i^* \sim N(0, 1)$, and a design matrix of $n = \{25, 50, 100, 250\}$ samples was generated from a multivariate normal distribution with a mean of zero, and Toeplitz correlation structure such that $\text{E}\left[x_{i,j} x_{i,k}\right] = \rho^{|j-k|}$, where $\rho = \{0.1, 0.5, 0.9\}$. Targets of the chosen distribution (normal, binomial, Poisson) were then generated using the regression coefficients $\boldsymbol{\beta}^*$ and generated design matrix. Maximum likelihood, MML and AIC_c were used to estimate the regression coefficients from the data, with $\boldsymbol{\Sigma} = \mathbf{I}_k$, and Kullback–Leibler (KL) divergences [12] from the true model were calculated for all three estimates.

The median ratios of the KL divergence obtained by the MML and AIC_c estimates over the KL divergence obtained by the maximum likelihood estimates are presented in Table 3. In all cases the ratio is less than or equal to one, and in many cases is substantially smaller than one, indicating that ridge regression offers an excellent alternative to maximum likelihood estimation. The improvements are generally larger for higher levels of correlation, which is expected given the nature of ridge regularisation. The MML estimates are competetive with, or superior to, the AIC_c estimates in all cases, and in the case of normal regression models MML is superior in all but one case.

6.2 Model Selection Experiments on Real Data

The MML and AIC_c ridge procedures were also tested in terms of model selection on several real datasets. Three datasets were chosen (two from the UCI machine learning repository [13], and one previously analysed in [14]): (i) the Pima indians dataset (binary targets, $q = 8$ covariates, $n = 768$ samples); (ii)

	Pima Indians		Diabetes		Boston Housing	
n	AIC_c	MML	AIC_c	MML	AIC_c	MML
25	3.775	1.063	1.659	1.404	5.991	3.895
50	1.552	0.641	1.281	1.233	3.692	3.370
100	0.542	0.528	1.151	1.144	3.200	3.181
250	0.501	0.501	1.101	1.103	3.057	3.099

Table 4. Kullback–Leibler divergences for three real datasets estimated by cross-validation.

the diabetes data (normal targets, $q = 10$, $n = 442$); and (iii) the Boston housing data (normal targets, $q = 14$, $n = 506$). Each dataset was randomly split into training and testing samples, and to make the task more difficult, four extra noise covariates generated from a standard normal distribution were appended to each training sample. MML and AIC_c were used to select a subset of the candidate regressors based on the training sample, with the potential subsets being determined from the path generated by the Lasso procedure [15]. The testing sample was subsequently used to assess the predictive performance of the criteria, measured in terms of mean KL divergence. Each test was repeated 100 times. The results are presented in Table 4, and show that MML is competetive with, or superior to, AIC_c for all three datasets, and for all sample sizes. The performance difference is especially noticable for the Pima indians dataset.

References

1. Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. Journal of the Royal Statistical Society. Series A (General) **135**(3) (1972) 370–384
2. Hoerl, A., Kennard, R.: Ridge regression. In: Encyclopedia of Statistical Sciences. Volume 8. Wiley, New York (1988) 129–136
3. Wallace, C.S., Boulton, D.M.: An information measure for classification. Computer Journal **11**(2) (August 1968) 185–194
4. 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
5. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. First edn. Information Science and Statistics. Springer (2005)
6. Farr, G.E., Wallace, C.S.: The complexity of strict minimum message length inference. Computer Journal **45**(3) (2002) 285–292
7. Makalic, E., Schmidt, D.F.: Minimum message length shrinkage estimation. Statistics & Probability Letters **79**(9) (2009) 1155–1161
8. Schmidt, D., Makalic, E.: MML invariant linear regression. In: Proceedings of the 22nd Australasian Joint Conference on Artificial Intelligence, Melbourne, Australia (2009) 312–321
9. Makalic, E., Schmidt, D.F.: MML logistic regression with translation and rotation invariant priors. In: Proceedings of the 25th Australasian Joint Conference on Artificial Intelligence, Sydney, Australia (2012)
10. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Second edn. Chapman & Hall/CRC (1989)
11. McQuarrie, A.D.R., Tsai, C.L.: Regression and Time Series Model Selection. World Scientific (1998)

12. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1) (March 1951) 79–86
13. Asuncion, A., Newman, D.: UCI machine learning repository. (2007)
14. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. The Annals of Statistics **32**(2) (April 2004) 407–451
15. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society (Series B) **58**(1) (1996) 267–288