Minimum Description Length Model Selection Criteria for Generalized Linear Models

MARK HANSEN
Bell Laboratories, Lucent Technologies

BIN YU
University of California, Berkeley

Abstract

This paper derives several model selection criteria for generalized linear models (GLMs) following the principle of Minimum Description Length (MDL). We focus our attention on the mixture form of MDL. Normal or normal-inverse gamma distributions are used to construct the mixtures, depending on whether or not we choose to account for possible over-dispersion in the data. For the latter, we use Efron's (1986) double exponential family characterization of GLMs as the likelihood. Standard Laplace approximations are employed to derive computationally tractable selection rules. Each of the criteria we construct have adaptive penalties on model complexity, either explicitly or implicitly. Theoretical results for the normal linear model, and a set of simulations for logistic regression illustrate that mixture MDL can "bridge" AIC and BIC in the sense that it can mimic the performance of either criterion, depending on which is best for the situation at hand.

KEY WORDS: AIC; Bayesian Methods; BIC; Code Length; Information Theory; Minimum Description Length; Model Selection; Generalized Linear Models.

1 Introduction

Statistical model selection attempts to decide between competing model classes for a data set. It is well-known that the maximum likelihood principle chooses the largest model under consideration, which in turn often overfits the data and has a poor prediction power. Many remedies have been proposed since the 70's to mark the start of model selection as a field. The most famous and widely used criteria are *A Information Criterion* (AIC) of Akaike (1973, 1977) and the *Bayesian Information Criterion* (BIC) of Schwarz (1978). They both take the form of a penalized maximized likelihood, but with different penalties: AIC has a fixed penalty of 1 for each additional dimension while BIC has $\log n/2$ (n is the sample size). The practical question is which one to use for one's problem. Theoretical and simulation studies (cf. Shibata, 1981, Speed and Yu, 1994, and references therein), mostly in the regression case, have revealed that depending on whether the underlying model is finite-dimensional or infinite-dimensional, one is preferred over the other. Since we never know the underlying model in practice, it is desirable to have model selection criteria which perform well independent of the underlying model. That is, we seek those criteria which adapt automatically to the situation at hand. In this paper, we derive such adaptive model selection criteria for generalized linear models under the Minimum Description Length (MDL) framework. Within MDL we find several generic prescriptions for constructing such selection criteria. In this paper, we focus on one MDL form that uses a so-called mixture code.

The MDL approach began with Kolmogorov's theory of algorithmic complexity, matured in the literature on information theory, and has recently received renewed interest within the statistics community. By viewing statistical modeling as a means of generating *descriptions* of observed data, the MDL framework (cf. Rissanen, 1989, Barron et al, 1998, and Hansen and Yu, 2001) discriminates between competing model classes based on the *complexity* of each description based on a model class. Precisely, the Minimum Description Length (MDL) Principle states that

Choose the model that gives the shortest description of data.

The complexity of a description is measured by the code length for the data based on the model. In general, there is a correspondence between the length of a prefix code on a finite set $\mathcal{A}$ and the quantity $-\log_2 Q$ for a probability distribution Q on $\mathcal{A}$. An integer-valued function L corresponds to the code length of a binary prefix code if and only if it satisfies Kraft's inequality

$$\sum_{a \in \mathcal{A}} 2^{-L(a)} \leq 1, \tag{1}$$

see Cover and Thomas (1991) for a proof. Therefore, given a prefix code $\mathcal{C}$ on $\mathcal{A}$ with length function L, we can define a distribution on $\mathcal{A}$ as follows,

$$Q(a) = \frac{2^{-L(a)}}{\sum_{z \in \mathcal{A}} 2^{-L(z)}} \quad \text{for any } a \in \mathcal{A}.$$

Conversely, for any distribution Q on $\mathcal{A}$ and any $a \in \mathcal{A}$, we can find a prefix code with length function $L(a) = \lceil -\log_2 Q(a) \rceil$, the smallest integer greater than or equal to $-\log_2 Q(a)$. Obviously, this inequality holds when A is the n-product space of a finite set, or when the data is an n-tuple. If the n-tuple data is assumed to come from a stochastic process, then when it is ergodic and stationary, an entropy can be defined for the infinite process. In the iid case the entropy simplifies to the expected value of the negative log probability of the 1-tuple. The entropy serves as a lower bound on the expected code length of any code, and is achieved by the code corresponding to the truth distribution of the process. For n-tuple data, most often Kraft's inequality holds asymptotically as an equality when Q is a distribution over the n-tuple and we can ignore the truncation error in making $-\log_2 Q(\cdot)$ into an integer. Moreover, since the base 2 is not crucial, we may call $-\log Q(\cdot)$ the code length of the code corresponding to Q.

Keeping in mind the correspondence between a code and a probability distribution warranted by Kraft's inequality, the search for description to use based on a model becomes that of the most efficient code or a probability distribution of the data based on the same model. The efficiency is measured in terms of coding gains relative to the lower bound entropy of the data generating process. When the competing models are members of a parametric family, using MDL to select a model or estimate a parameter is equivalent to maximum likelihood estimation (when the cost of transmitting the parameter estimate is fixed). To compare model classes or carry out model selection between model classes, efficient codes need to represent well the whole model classes. Such codes, justified rigorously in many models including iid and time series, can take on different forms. The most well-known is the two-stage code, with the first part as the code (or description) length for the encoding of the MLE parameter, and the second the code (or description) length of the data based on the model with the MLE parameter. Hence this form is also a penalized likelihood and with a first approximation, is exactly the same as BIC. The other forms include predictive, mixture and normalized maximum likelihood (NML). The predictive makes the most sense when the data come in sequentially and has a close connection to prequential inference; the mixture will be described in more detail in the next section; and the NML form is new and evolving, and we know the description expressions only in a few special cases including binomial model and Gaussian linear regression (cf. Rissanen, 1996, Barron et al, 1998, Hansen and Yu, 2001).

The rest of the paper is organized as follows. Section 2 gives the details of a mixture code in the context of regression-type models. Section 3 covers the gMDL model selection criterion from Hansen and Yu (2001) in the variance known and unknown cases to prepare the reader for the new results in Section 4. The criterion when σ^2 is known appeared originally in George and Foster (2001) in the context of a Bayesian analysis. Section 3.2 contains a new theorem to show the bridging effect of the gMDL criterion between AIC and BIC in a normal linear regression model.

Section 4 derives the gMDL criteria based on the mixture form of MDL. Normal or normal-inverse gamma distributions are used to construct the mixtures, depending on whether or not we choose to account for possible over-dispersion in the data. When the dispersion parameter is known, the resulting criterion appeared first in Peterson (1986) in the context of a Bayesian analysis. To account for dispersion effects, we use Efron's (1986) double exponential family characterization of GLMs as the likelihood. Standard Laplace approximations are employed to derive computationally tractable selection rules. Each of the criteria we construct have adaptive penalties on model complexity, either explicitly or implicitly. The last section of the paper contains a set of simulations for logistic regression to illustrate that mixture MDL can "bridge" AIC and BIC in the sense that it can mimic the performance of either criterion, depending on which is best for the situation at hand. The performance measures include the probability of selecting the correct model and test-error based on a selected model. The latter is found to be much less sensitive to the model selection criterion than the former due to the robustness of 0-1 loss in classification. MDL based criteria nevertheless have merits due to the conciseness of the models chosen by them.

2 Mixture MDL

In this paper we will consider regression-type models; that is, we would like to characterize the dependence of a random variable $Y \in \mathcal{Y} \subset \mathbb{R}$ on a vector of potential covariates $(X_1, \ldots, X_K) \in \mathbb{R}^K$. We consider various parametric model classes (2) (or conditional densities), indexed by a 0-1 binary vector $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K)$, for Y that involve a subset of the covariates corresponding to 1's in the model index vector $\boldsymbol{\gamma}$. Generically, we will let $\mathcal{M}_{\boldsymbol{\gamma}}$ denote a simple model class of the form with dimension $k_{\boldsymbol{\gamma}} = \sum_{j=1}^K \gamma_j$:

$$\sum_{j:\gamma_j=1} \beta_j X_j \tag{2}$$

and let $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = (\beta_k)_{\{j:\gamma_j=1\}}$. To fit this relationship, our basic data are observations of the form $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK})$. In observational studies it makes sense to consider $\boldsymbol{X}_i$ as being random, whereas in designed experiments the values of the covariates are specified. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$ denote the vector of responses and let $\boldsymbol{X}_K$ be the $n \times K$ full design matrix, $[\boldsymbol{X}_K]_{ij} = X_{ij}$. By $\boldsymbol{X}_{\boldsymbol{\gamma}}$ we mean a submatrix of $\boldsymbol{X}$ consisting of those columns j for which $\gamma_j = 1$. We connect the data $\boldsymbol{y} = (y_1, ..., y_n)$ to the model (2) via the conditional density functions $f_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}}(\boldsymbol{y}|\boldsymbol{X}_{\boldsymbol{\gamma}})$ for some set of parameters $\boldsymbol{\theta}_{\boldsymbol{\gamma}} \in \Theta$. (Typically, $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ will include regression parameters $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and possibly a dispersion effect.) In order to assess the suitability of $\mathcal{M}_{\boldsymbol{\gamma}}$, we derive a description length for $\boldsymbol{Y}$ based on $\mathcal{M}_{\boldsymbol{\gamma}}$.

For simplicity, we now drop the subscript $\boldsymbol{\gamma}$ except in places where we feel the need to remind the reader. The reader should interpret the model class $\mathcal{M}$, its dimension k, the design matrix $\boldsymbol{X}$, and the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as all depending on some subset of the available predictors. We then judge the appropriateness of this model based on the so-called *mixture form* of MDL. As its name suggests, this criterion starts with a mixture distribution that combines all the members in the class $\mathcal{M}$

$$m(\boldsymbol{y}|\boldsymbol{X}) = \int f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{X}) w(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}, \quad \boldsymbol{y} \in \mathcal{Y}^n, \tag{3}$$

where w is a probability density function on $\boldsymbol{\theta}$. This integral has a closed form expression when $f_{\boldsymbol{\theta}}(\cdot|\boldsymbol{X})$ is an exponential family and w is a conjugate distribution.

If $\mathcal{Y}$ is a finite set of values, we can use the distribution (3) to directly form a *mixture code* for strings $\boldsymbol{y} \in \mathcal{Y}^n$. In this setting, we assume that both sender and receiver know about the covariates $\boldsymbol{X}$, and we only have to transmit $\boldsymbol{y}$. As an example, suppose each $\mathcal{Y} = \{0, 1\}$ so that $\boldsymbol{Y}$ is a binary string of length n. We use the model class $\mathcal{M}$ and the distribution (3) to construct a mixture code for all 2^n strings $\boldsymbol{y} \in \{0,1\}^n$. Using one of several standard algorithms, it is possible to find a code that has the length function $L(\boldsymbol{y}) = \lceil -\log_2 m(\boldsymbol{y}|\boldsymbol{X}) \rceil$ for all $\boldsymbol{y} \in \mathcal{Y}^n$. This means that the number of bits required to transmit any $\boldsymbol{y} \in \{0,1\}^n$ is at most $-\log_2 m(\boldsymbol{y}|\boldsymbol{X})+1$ bits. The MDL principle then distinguishes between candidate model classes based on the associated length function $L(\boldsymbol{Y})$, the number of bits required to transmit the observed data $\boldsymbol{Y}$. As mentioned earlier, we have chosen to use base e in the log for our derivations.

When $Y_i \in \mathcal{Y} \subset \mathbb{R}$, $i = 1, \ldots, n$, is a continuous response, we form an approximate length function by discretizing the set $\mathcal{Y}$. That is, given a precision δ we obtain the description length

$$-\log \int f(\boldsymbol{y}|\theta, \boldsymbol{X}) w(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta} + n \log \delta. \tag{4}$$

Assuming that the precision used for this approximation is the same regardless of model class $\mathcal{M}$, we again arrive at the expression

$$-\log \int f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}) w(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta} \tag{5}$$

as a suitable length function. In the next section, we present a brief review of mixture MDL for the simple linear model. A full derivation of these results can be found in Hansen and Yu (2001). We review them here because they are structurally similar to the criteria we will derive abstractly for the entire class of GLMs.

When choosing between two model classes, the mixture form of MDL (with fixed hyperparameters) is equivalent to a Bayes factor (Kass and Raftery, 1995) based on the same distributions on the parameters spaces. As we will see in the next section, MDL allows for a natural, principled mechanism for dealing with hyperparameters that distinguishes it from classical Bayesian analysis. Also, keep in mind that w is not introduced as a prior in the Bayesian sense, but rather as a device for creating a distribution for the data $\boldsymbol{Y}$ from $\mathcal{M}$. This distinction also allows us more freedom in choosing w, and has spawned a number of novel applications in the engineering literature.

3 Regression

We begin with the simplest GLM, namely the normal linear model $\mathcal{M}_{\boldsymbol{\gamma}}$:

$$Y_i = \sum_{j:\gamma_j=1} \beta_j X_{ij} + \epsilon_i \,. \tag{6}$$

where the ϵ_i are normally distributed with mean zero and variance σ^2. To remind the reader that our basic model classes will consist of various subsets of the predictor variables $(X_1, \ldots, X_K)$, we restored the $\boldsymbol{\gamma}$ notation in the above equation. For simplicity, however, from this point on, we will drop it and consider derivations with respect to a single model class, a single choice of γ. Technically, we do not need to assume that the relationship in (6) holds for some collection of predictors $\boldsymbol{X}_K$, but instead we will entertain model classes because they are capable of capturing the major features observed in the observed data string $\boldsymbol{Y}$. For comparison with more general GLMs later, we treat separately the case in which σ^2 is known and unknown. In the former case, the parameter vector $\boldsymbol{\theta}$ in the mixture (3) consists only of the coefficients $\boldsymbol{\beta}$; while in the latter, $\boldsymbol{\theta}$ involves both $\boldsymbol{\beta}$ and σ^2.

We review this material because relatively straightforward, direct analyses yield the MDL selection criteria. When we tackle the complete class of GLMs, things become more difficult, but the final forms are reminiscent of those derived in this section.

3.1 Known error variance σ^2

Here, we take $\boldsymbol{\theta} = \boldsymbol{\beta}$ and let $w(\boldsymbol{\beta}|\boldsymbol{X})$ be a normal distribution with mean zero and variance-covariance matrix $\sigma^2 V$. As $\mathcal{Y} = \mathbb{R}$, we have to appeal to the discretized form of MDL (4). By using a conjugate distribution, we are able to perform the integration in (5) exactly. This leads to a code length of the form

$$\begin{aligned} L(\boldsymbol{y}|V) &= -\log m(\boldsymbol{y}|\boldsymbol{X}, V) \\ &= \frac{1}{2}\log|V^{-1} + \boldsymbol{X}^t\boldsymbol{X}| + \frac{1}{2}\log|V| + \frac{1}{2\sigma^2}\left(\boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}(V^{-1} + \boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}\right) , \end{aligned} \tag{7}$$

where we have dropped terms that depend only on n. We have also made explicit the dependence of the length function on the variance-covariance matrix V. Clearly, we can simplify this expression by taking $V = c(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ so that

$$-\log m(\boldsymbol{y}|\boldsymbol{X}, c) = \frac{k}{2}\log(1+c) + \frac{1}{2\sigma^2}\left(\boldsymbol{y}^t\boldsymbol{y} - \frac{c}{1+c}FSS\right) , \tag{8}$$

where $FSS = \boldsymbol{y}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$ is the usual fitted sum of squares corresponding to the OLS estimate $\widehat{\boldsymbol{\beta}}$,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} .$$

This particular choice of distribution is often attributed to Zellner (1986) who christened it the g-prior. Because the mixture form reduces to a relatively simple expression, the g-prior has been used extensively to derive Bayesian model selection procedures for the normal linear model. Under this prior, it is not hard to show that the posterior mean of $\boldsymbol{\beta}$ is $\frac{c}{1+c}\widehat{\boldsymbol{\beta}}$.

In (8) we have highlighted the dependence of the mixture on the scaling parameter c. George and Foster (2001) studied various approaches to setting c, establishing that certain values lead to well-known selection criteria like AIC and BIC.[1] Ultimately, they propose an empirical Bayes approach, selecting an estimate $\hat{c}$ via maximum likelihood. Hansen and Yu (2001) take a similar approach to the hyperparameter c, but motivate it from a coding perspective. We review this approach here. Essentially, each choice of c produces a different mixture distribution and hence a different code. Therefore, to let c depend on the data, both sender and receiver need to agree on which value of c to use. Hansen and Yu (2001) take a two-stage approach to hyperparameters like c; that is, c is transmitted first and then once each side knows which code to use, the data are sent. Of course, communicating c in this way will add to the code length, a charge that we make explicit by writing

$$L(\boldsymbol{y}) = L(\boldsymbol{y}|c) + L(c) = -\log m(\boldsymbol{y}|\boldsymbol{X}, c) + L(c) . \tag{9}$$

Following Rissanen (1989), the cost $L(c)$ is taken to be $\frac{1}{2}\log n$.[2] Minimizing (9) with respect to c gives us

$$\hat{c} = \max\left(\frac{FSS}{k\sigma^2} - 1, 0\right) , \tag{10}$$

and substituting into (8) yields

$$\begin{cases} \frac{\boldsymbol{y}^t\boldsymbol{y} - FSS}{2\sigma^2} + \frac{k}{2}\left[1 + \log\left(\frac{FSS}{\sigma^2 k}\right)\right] + \frac{1}{2}\log n & \text{for } FSS > k\sigma^2 \\ \\ \frac{\boldsymbol{y}^t\boldsymbol{y}}{2\sigma^2} & \text{otherwise} \end{cases} . \tag{11}$$

When the minimizing value of c is zero, the prior on $\boldsymbol{\beta}$ becomes a point mass at zero, effectively producing the "null" model corresponding to all effects being zero. This accounts for the second case in the above expression. We should note that the extra $\frac{1}{2}\log n$ penalty is essential to guarantee consistency of the selection method when the null model is true. The Bayesian criteria of George and Foster (2001) are basically the same, but leave off this extra term.

[1] A similar calibration between Bayesian methods and well-known selection criteria can also be found in Smith and Spiegelhalter (1980).

[2] The cost $\frac{1}{2}\log n$ can be motivated as follows: for regular parametric families, an unknown parameter can be estimated at rate $1/\sqrt{n}$. Hence there is no need to code such a parameter with a precision finer than $1/\sqrt{n}$. Coding c with precision $1/\sqrt{n}$ gives a cost to the first order $-\log[1/\sqrt{n}] = \log n/2$.

3.2 Unknown error variance

We now consider the regression model (6) when σ^2 is unknown. While George and Foster (2001) advocate estimating σ^2 and then applying the form (11), we prefer to assign a distribution to σ^2 and incorporate it into the mixture. Following Hansen and Yu (2001), we employ a conjugate normal-inverse gamma distribution to form the mixture code; that is, $\frac{1}{\sigma^2}$ has a gamma distribution with shape parameter a; and given σ^2, $\boldsymbol{\beta}$ is normal with mean zero and variance $\sigma^2 V$. Setting $\tau = \sigma^2$, these densities are given by

$$w(\boldsymbol{\beta}, \tau) \;\propto\; \tau^{\frac{-(k+3)}{2}} \exp\left[\frac{-\boldsymbol{\beta}^t V^{-1}\boldsymbol{\beta} + a}{2\tau}\right], \tag{12}$$

where a and V are hyperparameters. Under this class of priors, the mixture distribution (3) has the form

$$\begin{aligned} -\log m(\boldsymbol{y}|\boldsymbol{X}, a, V) \;=\; & \frac{1}{2}\log|V^{-1} + \boldsymbol{X}^t\boldsymbol{X}| + \frac{1}{2}\log|V| \\ & -\frac{1}{2}\log a + \frac{n+1}{2}\log\left(a + \boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}(V^{-1} + \boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}\right), \end{aligned} \tag{13}$$

where we have ignored terms that do not depend on our particular choice of model. The derivation of $m(\boldsymbol{y}|\boldsymbol{X}, a, V)$, the marginal or predictive distribution of $\boldsymbol{y}$, is standard and can be found in O'Hagan (1994).

Our approach to handling the hyperparameter a will be the same as that in the previous section. Minimizing (13) with respect to a we find that $\hat{a} = (\boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}(V^{-1} + \boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y})/n$ which leaves

$$-\log m(\boldsymbol{y}|\boldsymbol{X}, \hat{a}, V) = \frac{1}{2}\log|V| + \frac{1}{2}\log|V^{-1} + \boldsymbol{X}^t\boldsymbol{X}| + \frac{n}{2}\log\left(\boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}(V^{-1} + \boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}\right), \tag{14}$$

As was true in the known-variance case, we can achieve a simplification in computing the mixture distribution if we again make Zellner's choice of $V = c(\boldsymbol{X}^t\boldsymbol{X})^{-1}$. This leaves

$$-\log m(\boldsymbol{y}|\boldsymbol{X}, \hat{a}, c) = \frac{k}{2}\log(1+c) + \frac{n}{2}\log\left(\boldsymbol{y}^t\boldsymbol{y} - \frac{c}{1+c}FSS\right), \tag{15}$$

To settle the hyperparameter c, we again minimize the overall code length to find

$$\hat{c} = \max(F - 1, 0) \quad \text{with} \quad F = \frac{FSS}{kS}, \tag{16}$$

where F is the usual F-ratio for testing the hypothesis that each element of $\boldsymbol{\beta}$ is zero, and $S = RSS/(n-k)$. The truncation at zero in (16) rules out negative values of the prior variance. Rewriting (16), we find that $\hat{c}$ is zero unless $R^2 > k/n$, where R^2 is the usual squared multiple correlation coefficient. When the value of $\hat{c}$ is zero, the prior on $\boldsymbol{\beta}$ becomes a point mass at zero, effectively producing the "null" mixture model[3] corresponding to zero regression effects. Substituting the optimal value of $\hat{c}$ into (15) and adding the cost to code the hyperparameters a and c as we did before, we arrive at a final mixture form

$$\text{gMDL} \;=\; \begin{cases} \frac{n}{2}\log S + \frac{k}{2}\log F + \log n\,, & R^2 \geq k/n \\ \frac{n}{2}\log\left(\frac{y'y}{n}\right) + \frac{1}{2}\log n\,, & \text{otherwise}\,. \end{cases} \tag{17}$$

3.3 Comparison

As alluded to in the introduction, two widely used model selection criteria are AIC and BIC. In the case of regression with an unknown variance, they take forms

$$\text{AIC} = \frac{n}{2}\log RSS + k \quad \text{and} \quad \text{BIC} = \frac{n}{2}\log RSS + \frac{k}{2}\log n\,. \tag{18}$$

[3] The null model is a scale mixture of normals, each $N(0,\tau)$ and τ having an inverse-gamma prior.

Comparing these with (17), we see that the essential difference is in the penalty. Both AIC and BIC have non-data dependent penalties, while gMDL has a data-dependent $\log F/2$ for each additional dimension.

By charging less for each new variable, AIC tends to include more terms. When the underlying model consists of many effects, or more precisely the model is infinite-dimensional, AIC tends to perform better. If we take our figure of merit to be prediction error, then AIC has been shown to be optimal in this setting both through theoretical and simulation studies. When the true, data generating mechanism is finite-dimensional (and is included among the candidates we are comparing), the stronger penalty of BIC tends to perform better. For this kind of problem, we can also judge selection criteria based on consistency (which leads to prediction optimality), that is whether or not they ultimately select the correct model as the number of samples tends to infinity, and BIC has been shown to perform optimally.

We will now demonstrate that gMDL with its adaptive penalty enjoys the advantages of both AIC and BIC in the regression context. We focus on the simple linear model because the expressions are easy to work with, although we expect the same kind of result will hold for GLMs. To simplify our analysis, we assume the regressors are ordered as $x_{i1}, x_{i2}, \ldots$. Following Breiman and Freedman (1983), we assume that $X_i = (X_{i1}, X_{i2}, ..., X_{ij}, ...)$ are Gaussian, zero-mean random vectors and let

$$\sigma_k^2 = Var[\sum_{j=k+1}^{\infty} \beta_j X_{ij} | X_{i1}, \ldots, X_{ik}].$$

Then the finite-dimensional model assumption implies that $\sigma_{k_0}^2 = 0$ for some $k_0 > 0$. Using similar arguments as those used to prove Theorem 1.4 in Breiman and Freedman (1983) and the fact that $||y|| = \sigma^2(1 + o_p(1))$, it is easy to show that

Theorem: $F = [\frac{n}{k}\frac{\sigma_0^2 - \sigma_k^2}{\sigma_k^2 + \sigma^2} + 1](1 + o_p(1))$ where $o_p(1) \to 0$ in probability uniformly over $0 \le k \le n/2$.

Corollary: If the model is finite-dimensional and the maximum dimension examined $K = K_n = o(n)$, gMDL is consistent and is also prediction-optimal.

The above theorem gives an expansion on the data dependent-penalty of gMDL, and the corollary establishes the optimality of gMDL as those enjoyed by BIC when the model is finite-dimensional. When $\sigma_k^2 > 0$ for all k the model is infinite-dimensional or nonparametric. The factor $(\sigma_0^2 - \sigma_k^2)/k(\sigma_k^2 + \sigma^2)$ can be viewed as the average signal to noise ratio for the fitted model. Adjusting the penalty with a half of log of this factor, gMDL is able to adapt to perform well in terms of prediction in both domains, parametric or nonparametric, where one of either AIC or BIC is known to be optimal. The simulation studies in Hansen and Yu (2001) support this adaptivity of gMDL since there gMDL has an overall prediction performance better than AIC or BIC.

In the next section, we will show that the newly derived MDL-based criteria for GLMs possess the same adaptivity to AIC or BIC as gMDL does.

4 Generalized Linear Models

The characterization of a GLM starts with an exponential family of the form

$$f(y) = \exp\left(\frac{y\psi - b_1(\psi)}{b_2(\phi)} + b_3(y, \phi)\right), \quad y \in \mathcal{Y}, \tag{19}$$

where b_1, b_2 and b_3 are known functions. We refer to ψ as the canonical parameter for the family. Typically, we take $b_2(\phi) = \phi$, and refer to ϕ as the dispersion parameter. It plays the role of the noise-variance in the ordinary regression setup of the previous section. The family (19) contains many practically important cases, including the normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions. With this model, it is not hard to show that if Y has distribution (19),

$$\begin{array}{rcccl} E(Y) & = & \mu & = & b_1'(\psi) \\ \mathrm{var}(Y) & = & \sigma^2 & = & b_1''(\psi) b_3(\phi) \end{array} . \tag{20}$$

As with the normal case above, the GLM framework allows us to study the dependence of a response variable $Y \in \mathcal{Y}$ on a vector of covariates $(X_1, \ldots, X_K)$. Each model class corresponds to some value of the binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$, and we relate the mean μ of Y to a subset of the covariates via the linear predictor

$$\eta = g(\mu) , \qquad \text{for } \eta = \sum_{j:\boldsymbol{\gamma}_j = 1} \beta_j X_j , \tag{21}$$

where g is a one-to-one, continuously differentiable transformation known as the link function. Using (20) and (21) we see that $\eta = g(b'(\psi))$.[4] Again we let $\boldsymbol{\beta}_\gamma = (\beta_j)_{j:\gamma_j=1}$ denote the vector of regression coefficients and $k_\gamma = \sum \gamma_j$ its dimension. The unknown parameters associated with this model are denoted $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ and include both $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ as well as a possible dispersion effect ϕ. We observe data of the form $(Y_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, n$ where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK})$ and again $\boldsymbol{X}_K$ is the $n \times K$ full design matrix $[\boldsymbol{X}_K]_{ij} = X_{ij}$. We let $\boldsymbol{X}_{\boldsymbol{\gamma}}$ refer to a submatrix of $\boldsymbol{X}_K$ consisting of only those columns j for which $\gamma_j = 1$. Let $f_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}}(\boldsymbol{y}|\boldsymbol{X}_{\boldsymbol{\gamma}})$ denote the density for $\boldsymbol{Y}$ based on model class $\boldsymbol{\gamma}$.

As with our treatment of the regression context, maintaining the model index $\boldsymbol{\gamma}$ needlessly complicates our derivations. From this point on, we again drop it, reminding the reader that terms like $\mathcal{M}$, $\boldsymbol{X}$, k and $\boldsymbol{\beta}$ all refer to a specific subset of covariates. For all the GLM cases, we will begin with a Laplace approximation to the mixture form which will be exact for the normal linear model. That is, we start with

$$m(\boldsymbol{y}|\boldsymbol{X}) \approx (2\pi)^{\frac{k}{2}} | - H^{-1}(\tilde{\boldsymbol{\beta}})|^{\frac{1}{2}} f_{\tilde{\boldsymbol{\beta}}}(\boldsymbol{y}|\boldsymbol{X}) w(\tilde{\boldsymbol{\beta}}) \tag{22}$$

where H is the Hessian of $h(\boldsymbol{\beta}) = \log f_{\boldsymbol{\beta}}(\boldsymbol{y}|\boldsymbol{X}) + \log w(\boldsymbol{\beta})$ and $\tilde{\boldsymbol{\beta}}$ is the posterior mode of $\boldsymbol{\beta}$. In working with this form, we will repeatedly make use of the Fisher information matrix

$$I(\boldsymbol{\beta}) = \boldsymbol{X}^t \boldsymbol{W}(\boldsymbol{\beta}) \boldsymbol{X} , \tag{23}$$

where $\boldsymbol{W}$ is a diagonal weight matrix. Note that for GLMs, the observed Fisher information is the same as the Fisher information.

Form (22) is still difficult to work with in practice because there is typically no closed-form expression for the posterior mode. We will now consider several criteria that make sensible choices for f and w that lead to computationally tractable criteria.

4.1 Direct approach

In this section, we will derive a criterion that first appeared in Peterson (1986). As with the regression context, the original motivation for this form was not MDL, but rather an approximation to a full Bayesian approach. Our analysis will follow closely the case of σ^2 known for regression. Let $\boldsymbol{\beta}$ be the MLE of $\boldsymbol{\beta}$, and assume that the prior $w(\boldsymbol{\beta})$ is normal with mean zero and variance-covariance V. Then, we can approximate $\tilde{\boldsymbol{\beta}}$ via a single Newton step

$$\begin{array}{rcl} \tilde{\boldsymbol{\beta}} & \approx & \boldsymbol{\beta} - H(\boldsymbol{\beta})^{-1} h'(\boldsymbol{\beta}) \\ & \approx & \boldsymbol{\beta} + (I(\boldsymbol{\beta}) + V^{-1})^{-1} V \boldsymbol{\beta} \end{array}$$

[4] Taking $b' = g^{-1}$ means that the canonical parameter ψ and the linear predictor η are the same. This choice of g is known as the canonical link function.

using the fact that $H(\boldsymbol{\beta}) = -(I(\boldsymbol{\beta}) + V^{-1})$, where $I(\widehat{\boldsymbol{\beta}})$ is the Fisher information evaluated at $\widehat{\boldsymbol{\beta}}$. We now focus on the case where the prior variance-covariance matrix for $\boldsymbol{\beta}$ is simply $cI(\widehat{\boldsymbol{\beta}})^{-1}$. For the normal linear model, this leads us to Zellner's g-prior. Unfortunately, for the other important members of this family, the prior variance-covariance matrix will depend on $\widehat{\boldsymbol{\beta}}$. From a strict coding perspective this is hard to accept; it would imply that sender and receiver both know the coefficient $\widehat{\boldsymbol{\beta}}$. Nonetheless, it is instructive to follow this line of analysis and compare it with the results of the previous section. For $V = cI(\widehat{\boldsymbol{\beta}})^{-1}$ we find that the one-step Newton-Raphson iteration gives

$$\tilde{\boldsymbol{\beta}} \approx \frac{c}{1+c}\widehat{\boldsymbol{\beta}}$$

which agrees with our regression form of MDL when σ^2 is known.

Continuing with the expression (22), we find that

$$\begin{aligned} \log w(\tilde{\boldsymbol{\beta}}) &\approx \log w\left(\frac{c}{1+c}\widehat{\boldsymbol{\beta}}\right) \\ &= -\frac{k}{2}\log 2\pi + \log|cI(\widehat{\boldsymbol{\beta}})| - \frac{1}{2}\frac{c}{(1+c)^2}\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} \end{aligned} \tag{24}$$

and that after a Taylor expansion of $\log f_{\boldsymbol{\beta}}(\boldsymbol{y}|\boldsymbol{X})$ around $\widehat{\boldsymbol{\beta}}$

$$\begin{aligned} \log f_{\tilde{\boldsymbol{\beta}}}(\boldsymbol{y}|\boldsymbol{X}) &\approx \log f_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{y}|\boldsymbol{X}) - \frac{1}{2}(\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})^t I(\widehat{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}) \\ &= \log f_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{y}|\boldsymbol{X}) - \frac{1}{2}\frac{1}{(1+c)^2}\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} \end{aligned} \tag{25}$$

Combining (24) and (25) we arrive at the expression

$$\log f_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{y}|\boldsymbol{X}) + \log w(\tilde{\boldsymbol{\beta}}) \approx \log f_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{y}|\boldsymbol{X}) - \frac{1}{2}\frac{1}{1+c}\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} - \frac{k}{2}\log 2\pi + \frac{k}{2}\log c + \frac{1}{2}\log|I(\widehat{\boldsymbol{\beta}})|\,. \tag{26}$$

Finally, collecting terms in (22) we find an expression for the code length given c

$$-\log m(\boldsymbol{y}|\boldsymbol{X}) \approx \frac{k}{2}\log(1+c) + \frac{1}{2}\frac{1}{1+c}\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} - \log f_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{y}|\boldsymbol{X})\,.$$

We then eliminate the hyperparameter c using the same minimization approach in (9). This yields

$$\hat{c} = \max\left(\frac{\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}}{k} - 1, 0\right).$$

Substituting this in the mixture form, we find the final MDL criterion

$$\begin{cases} -\log f_{\widehat{\boldsymbol{\beta}}}(y|X,\widehat{\boldsymbol{\beta}}) + \frac{k}{2}\left[1 + \log\left(\frac{\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}}{k}\right)\right] + \frac{1}{2}\log n & \text{for } \widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} > k \\ -\log f_{\mathbf{0}}(y|X) & \text{otherwise} \end{cases}$$

The function $f_{\mathbf{0}}(y|X)$ represents the log-likelihood when all the regression effects are zero. Again, we have added an extra $\frac{1}{2}\log n$ term to the top expression to account for the coding cost of c. This corresponds exactly to the regression context when σ^2 is known.

4.2 Accounting for over-dispersion

In many families, like the Poisson and binomial models, the dispersion parameter is fixed $\phi = 1$. However, in practice it is often the case that the data do not support this, forcing us to consider over-dispersed models. There are a variety of ways to introduce extra variability into the form (19), many of which are primarily meant as computational devices. Efron (1986) constructs a family to explicitly account for over-dispersion that admits an analysis for GLMs similar to that for ordinary regression in the σ^2-unknown case. A related technique was independently derived by West (1985).

To understand this form, we have to first rewrite the log-likelihood for a GLM in terms of its mean vector $l(\boldsymbol{y}|\boldsymbol{\mu})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Now, using this notation, without the restriction (21) on the mean, the maximum value of the log-likelihood is simply $l(\boldsymbol{y}|\boldsymbol{y})$. We then define the deviance as the difference

$$D(\boldsymbol{y}|\boldsymbol{\beta}) = 2l(\boldsymbol{y}|\boldsymbol{y}) - 2l(\boldsymbol{y}|\boldsymbol{\mu}) \, ,$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients that yield $\boldsymbol{\mu}$ through (21). To incorporate a dispersion parameter, Efron (1986) motivates the use of

$$\tau^{-n/2} e^{l(\boldsymbol{y}|\boldsymbol{\mu})/\tau + (1-1/\tau) l(\boldsymbol{y}|\boldsymbol{y})} \tag{27}$$

as an (approximate) likelihood. Technically, this expression should include a normalizing constant $C(\tau, \boldsymbol{\beta})$. Following Efron (1986), however, it can be shown that $C(\tau, \boldsymbol{\beta}) = 1 + O(n^{-1})$, and hence can be ignored for reasonable sample sizes. Rewriting (27), we will work with

$$\tau^{-n/2} e^{-\frac{(2l(\boldsymbol{y}|\boldsymbol{y}) - 2l(\boldsymbol{y}|\boldsymbol{\mu}))}{2\tau}} e^{l(\boldsymbol{y}|\boldsymbol{y})} = \tau^{-n/2} e^{-\frac{D(\boldsymbol{y}|\boldsymbol{\beta})}{2\tau}} e^{l(\boldsymbol{y}|\boldsymbol{y})} \; , . \tag{28}$$

Then, arguing as we did for the σ^2 unknown case in regression, we use a normal-inverse gamma prior with variance-covariance matrix τV. The joint probability of $\boldsymbol{\beta}$, τ and $\boldsymbol{y}$ is given by

$$\tau^{-1-(n+1)/2} \frac{e^{-a/2\tau}}{\sqrt{\pi}} \sqrt{\frac{a}{2}} (2\pi\tau)^{-k/2} |V^{-1}|^{1/2} e^{\frac{-\boldsymbol{\beta}^t V^{-1} \boldsymbol{\beta} - D(\boldsymbol{y}|\boldsymbol{\beta})}{2\tau}} \tag{29}$$

To integrate out $\boldsymbol{\beta}$, we use the Laplace method again which this time yields

$$\frac{\tau^{-1-(n+1)/2}}{\sqrt{\pi}} \sqrt{\frac{a}{2}} |V^{-1}|^{1/2} |V^{-1} + I(\widehat{\boldsymbol{\beta}})|^{-1/2} e^{\frac{-a - \tilde{\boldsymbol{\beta}}^t V^{-1} \tilde{\boldsymbol{\beta}} - D(\boldsymbol{y}|\tilde{\boldsymbol{\beta}})}{2\tau}} \tag{30}$$

where $I(\tilde{\boldsymbol{\beta}})$ is the Fisher information matrix evaluated at the posterior mode $\tilde{\boldsymbol{\beta}}$. Integrating with respect to τ then yields

$$-\log m(\boldsymbol{y}|a, V) = \frac{n+1}{2} \log \left(a + \tilde{\boldsymbol{\beta}}^t V^{-1} \tilde{\boldsymbol{\beta}} + D(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}) \right) - \frac{1}{2} \log a + \frac{1}{2} \log |V| + \frac{1}{2} \log |V^{-1} + I(\tilde{\boldsymbol{\beta}})| \tag{31}$$

Following the prescription in the regression context, we eliminate the hyperparameter a by minimizing the overall code length. In this case we easily find that

$$-\log m(\boldsymbol{y}|\hat{a}, V) = \frac{n}{2} \log \left(\tilde{\boldsymbol{\beta}}^t V^{-1} \tilde{\boldsymbol{\beta}} + D(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}) \right) + \frac{1}{2} \log |V| + \frac{1}{2} \log |V^{-1} + I(\tilde{\boldsymbol{\beta}})| \tag{32}$$

We have now obtained a usable criterion for model selection. Specifying V, we can compute $\tilde{\boldsymbol{\beta}}$ with simple Newton-Raphson iterations. In our regression analysis, we used Zellner's g-prior for $\boldsymbol{\beta}$ which led to a closed-form selection criterion. The analog in this case is $V = cI^{-1}(\widehat{\boldsymbol{\beta}})$. In the case of a GLM, this choice is somewhat unsettling because $I(\widehat{\boldsymbol{\beta}})$ is computed at the MLE. If we were to adhere to a strict MDL setting, it would not make sense; from a coding perspective, both sender and receiver would have to know about $\widehat{\boldsymbol{\beta}}$. Recall that for a GLM, the Fisher information matrix takes the form $\boldsymbol{X}^t\boldsymbol{W}(\widehat{\boldsymbol{\beta}})\boldsymbol{X}$ where W is a diagonal weight matrix. One simple alternative is to take $V = c(\boldsymbol{X}^t\boldsymbol{X})^{-1}$, or $V = c\mathbf{1}$, where $\mathbf{1}$ is the identity matrix. In each of these cases, we must either approximate the $\tilde{\boldsymbol{\beta}}$ or iterate to find it. We will consider both kinds of selection criteria.

Following the approximation route, if we choose $V = cI^{-1}(\widehat{\boldsymbol{\beta}})$, we get

$$\tilde{\boldsymbol{\beta}} \approx \frac{c}{1+c}\widehat{\boldsymbol{\beta}} \tag{33}$$

and

$$\frac{k}{2}\log(1+c) + \frac{n}{2}\log\left(\frac{1}{1+c}\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} + D(\boldsymbol{y}|\widehat{\boldsymbol{\beta}})\right) \tag{34}$$

Here we have substituted in the one-step Newton-Raphson approximation for $\tilde{\boldsymbol{\beta}}$ and have approximated the deviance $D(\boldsymbol{y}|\tilde{\boldsymbol{\beta}})$ by a Taylor expansion around $\widehat{\boldsymbol{\beta}}$ and used a relation from Raftery (1996). Maximizing with respect to c yields

$$\hat{c} = \max\left(\frac{(n-k)\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}}{kD(\boldsymbol{y}|\widehat{\boldsymbol{\beta}})} - 1, 0\right) \tag{35}$$

This then gives the form

$$\begin{cases} \frac{n}{2}\log\frac{D(\boldsymbol{y}|\widehat{\boldsymbol{\beta}})}{n-k} + \frac{k}{2}\log\frac{\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}/k}{D(\boldsymbol{y}|\widehat{\boldsymbol{\beta}})/(n-k)} + \log n & \text{if } F > 1 \\ \frac{n}{2}\log\frac{D(\boldsymbol{y}|\mathbf{0})}{n} + \frac{1}{2}\log n & \text{otherwise} \end{cases}, \tag{36}$$

where $D(\boldsymbol{y}|\mathbf{0})$ represents the deviance calculated under a model with zero regression effects.

For the other choices of $V^{-1} = c\Sigma$, we do not have a closed-form expression for the maximizing c. Instead, we can perform a search, but this is best done in conjunction with finding $\tilde{\boldsymbol{\beta}}$. It is also possible to use our approximate $\tilde{\boldsymbol{\beta}}$ (33) to derive a simple iteration to find c. In this case, we find

$$c = \frac{kR_c}{n\widehat{\boldsymbol{\beta}}^t I(\widehat{\boldsymbol{\beta}})\left(I(\widehat{\boldsymbol{\beta}}) + c\Sigma\right)^{-1}\Sigma\left(I(\widehat{\boldsymbol{\beta}}) + c\Sigma\right)^{-1} I(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} + R_c\mathrm{trace}\left(c\mathbf{1} + I(\widehat{\boldsymbol{\beta}})\Sigma^{-1}\right)} \tag{37}$$

where

$$R_c = D(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}) + c\widehat{\boldsymbol{\beta}}^t\Sigma\widehat{\boldsymbol{\beta}} - c^2\widehat{\boldsymbol{\beta}}^t\Sigma(I(\widehat{\boldsymbol{\beta}}) + c\Sigma)^{-1}\Sigma\widehat{\boldsymbol{\beta}}\,. \tag{38}$$

Convergence of this algorithm is usually fairly fast, although as we will see, it can depend on the starting values.

5 Simulations

We have chosen 8 different simulation setups to compare AIC and BIC with the new MDL-based criteria derived in this section. We focus on logistic regression, and consider $K = 5$ potential covariates. We specify two distributions on $\boldsymbol{X}$. In the first, each column consists of 100 observations from a standard normal distribution and the different columns are independent. In the second case, we again use normal covariates, but now we consider a correlation structure of the form

$$\mathrm{cov}(X_i, X_j) = \rho^{|i-j|}, \text{ for } i,j = 1,\ldots,5.$$

In the simulations below, we took $\rho = 0.75$. The data $\boldsymbol{Y}$ was then generated using the standard logistic GLM using one of 8 different coefficient vectors. All $2^5 = 32$ possible models were fit and compared using the various selection criteria. In Table 1, we present the classification error rate for each procedure: Column 4 corresponds to mixture MDL with a normal-inverse gamma mixing distribution to capture dispersion effects and $V^{-1} = cI(\widehat{\boldsymbol{\beta}})$ (Section 4.2); Column 5 corresponds to mixture MDL with a fixed dispersion parameter ϕ and hence a normal mixing distribution again with $V^{-1} = cI(\widehat{\boldsymbol{\beta}})$ (Section 4.1); Columns 6 and 7 are BIC and AIC (18). Columns 8 through 11 also make use of the normal-inverse gamma distribution but with different choices of the variance-covariance matrix V^{-1}; $c\mathbf{1}$ for 8 and 10, and $c\boldsymbol{X}^t\boldsymbol{X}$ for 9 and 11. Columns 8 and 10 differ only in how we estimate $\tilde{\boldsymbol{\beta}}$ and $\hat{c}$; in the first case the iteration (37) is used, while in the second a full search is performed to identify both $\tilde{\boldsymbol{\beta}}$ and the appropriate value of c. The same holds for Columns 9 and 11, but with the different variance-covariance matrix.

In Table 1 we see that most of the selection criteria behave the same, at least in terms of classification error; this 0-1 error is very robust. In Tables 2, 3 and 4, we illustrate the types of models selected by each scheme. The column headings are of the form $x - y$ where x denotes the number of variables correctly included in the model and y denotes the number of excess variables. So, for Table 2, the true model $(2, 0, 0, 0, 0)$ consists of only one effect. The heading "1-0" represents the correct model, while the column "1-1" means that one extra term was included. From these tables, we see that the three MDL criteria (Columns 8, 10 and 11) adapt to either AIC or BIC depending on which performs better in all 8 set-ups. Column 9 seemed to have some problems, and we believe this is because the iterations (37) failed to converge properly (possibly due to the approximations used to generate generate the form). Finally, we see that the columns using $I(\widehat{\boldsymbol{\beta}})$ can perform poorly. Recall that we derived these forms even though their reliance on $\widehat{\boldsymbol{\beta}}$ violates the basic coding ideas behind MDL.

To consider the cases in more depth, we start with Table 2. Here "truth" is a small model, $(2, 0, 0, 0, 0)$, an ideal case for BIC. Clearly, this criterion selects the right model more often than the other procedures. The mixture MDL procedures that use variance-covariance matrices other than $I(\widehat{\boldsymbol{\beta}})$ also perform quite well. In terms of test error, each of the procedures are about the same. Overall, we can recommend the MDL based criteria in terms of their ability to adapt and select concise models.

In Table 3, the coefficient vector is $(3, 2, 2, 0, 0)$, a middle-ground case, in which BIC is the best. The $I(\widehat{\boldsymbol{\beta}})$ criteria perform rather poorly, as does the $\boldsymbol{X}^t\boldsymbol{X}$ case with iterations (37) to find $\hat{c}$. In the latter case, the poor performance is even reflected in the prediction error. Again, we intend to examine whether it is the approximation that led to (37) that caused the problem, or if it was poor starting values for the iterations.

Finally, in Table 4, we consider a large model with coefficient vector $(5, 1, 1, 1, 1)$, an ideal situation for AIC. Here we see that BIC fails to capture the correct model form, and the test error is slightly worse as a result. All the MDL criteria outperform even AIC in terms of identifying the correct model, although this does not translate into significant test error improvements.

6 Acknowledgements

B. Yu's research is partially supported by grants from NSF (FD01-12731) and ARO (DAAD19-01-1-0643). M. Hansen would like to thank John Chambers, Diane Lambert, Daryl Pregibon and Duncan Temple Lang for helpful discussions.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. AC*, **19**, 716-723.

[2] Akaike, H. (1977). An objective use of Bayesian models. *Ann. Inst. Statist. Math.*, **29**, 9 -20.

Coefficients	ρ	Bayes Rate	Mix ϕ	$\phi = 1$	BIC	AIC	$\mathbf{1}$ Iter	$\boldsymbol{X}^t\boldsymbol{X}$ Iter	$\mathbf{1}$ search	$\boldsymbol{X}^t\boldsymbol{X}$ search
3 2 2 0 0	0	0.125	0.138	0.138	0.135	0.137	0.137	0.138	0.137	0.137
	0.75	0.087	0.101	0.101	0.104	0.101	0.100	0.104	0.100	0.101
5 1 1 1 1	0	0.098	0.115	0.116	0.128	0.118	0.120	0.118	0.118	0.118
	0.75	0.072	0.087	0.087	0.092	0.089	0.087	0.087	0.087	0.089
2 2 2 2 2	0	0.115	0.130	0.130	0.131	0.130	0.130	0.131	0.130	0.130
	0.75	0.067	0.081	0.083	0.095	0.086	0.081	0.081	0.081	0.087
3.0 1.5 1.5 0.5 0.0	0	0.137	0.153	0.152	0.154	0.152	0.153	0.155	0.153	0.153
	0.75	0.093	0.108	0.108	0.112	0.109	0.108	0.111	0.108	0.109
5 0 0 0 0	0	0.103	0.116	0.114	0.110	0.114	0.113	0.113	0.113	0.113
	0.75	0.104	0.116	0.115	0.109	0.114	0.114	0.112	0.114	0.113
2 0 0 0 0	0	0.220	0.233	0.233	0.228	0.233	0.230	0.230	0.230	0.230
	0.75	0.221	0.231	0.231	0.227	0.232	0.230	0.229	0.230	0.229
1.5 1.5 1.5 1.5 0.0	0	0.163	0.177	0.177	0.177	0.177	0.177	0.180	0.177	0.177
	0.75	0.101	0.119	0.120	0.129	0.121	0.118	0.124	0.118	0.122
8 4 2 1 0	0	0.060	0.074	0.074	0.077	0.074	0.075	0.074	0.074	0.074
	0.75	0.040	0.057	0.058	0.060	0.058	0.057	0.057	0.057	0.058

Table 1: Classification errors for the different selection criteria.

Criterion	0-1	0-2	0-3	0-4	1-0	1-1	1-2	1-3	1-4
Mix ϕ	0	0	0	0	131	70	37	12	0
$\phi = 1$	0	0	0	0	134	72	37	7	0
BIC	0	0	0	0	215	31	4	0	0
AIC	0	0	0	0	121	85	40	4	0
$\mathbf{1}$, iter	0	0	0	0	176	55	17	2	0
$\boldsymbol{X}^t\boldsymbol{X}$, iter	0	0	0	0	179	53	18	0	0
$\mathbf{1}$, search	0	0	0	0	183	51	15	1	0
$\boldsymbol{X}^t\boldsymbol{X}$, search	0	0	0	0	179	51	19	1	0

Table 2: Coefficient: (2, 0, 0, 0, 0). Number of times different sized models were selected.

Criterion	0-1	0-2	1-0	1-1	1-2	2-0	2-1	2-2	3-0	3-1	3-2
Mix ϕ	0	0	0	0	0	0	0	0	111	103	36
$\phi = 1$	0	0	0	0	0	0	0	0	134	93	23
BIC	0	0	0	0	0	1	0	0	227	20	2
AIC	0	0	0	0	0	0	0	0	173	71	6
$\mathbf{1}$, iter	0	0	0	0	0	0	0	0	176	49	25
$\boldsymbol{X}^t\boldsymbol{X}$, iter	0	0	0	0	0	3	0	0	71	31	145
$\mathbf{1}$, search	0	0	0	0	0	0	0	0	184	61	5
$\boldsymbol{X}^t\boldsymbol{X}$, search	0	0	0	0	0	0	0	0	180	64	6

Table 3: Coefficient: (3, 2, 2, 0, 0). Number of times different sized models were selected.

Criterion	1-0	2-0	3-0	4-0	5-0
Mix ϕ	0	0	9	56	185
$\phi = 1$	0	1	12	70	167
BIC	4	35	64	89	58
AIC	0	2	24	94	130
$\mathbf{1}$, iter	0	4	33	79	134
$\boldsymbol{X}^t\boldsymbol{X}$, iter	1	19	13	18	199
$\mathbf{1}$, search	0	3	23	86	138
$\boldsymbol{X}^t\boldsymbol{X}$, search	0	3	23	89	135

Table 4: Coefficient: (5, 1, 1, 1, 1). Number of times different sized models were selected.

[3] Barron, A., Rissanen, J., and Yu, B. (1998). Minimum Description Length principle in coding and modeling. *IEEE Trans. Inform. Theory.* (Special Commemorative Issue: Information Theory: 1948-1998), **44**, 2743-2760

[4] Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* , **78**, 131-136.

[5] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* New York: John Wiley & Sons.

[6] Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* , **81**, 709-721.

[7] George, E. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, **88**, 881–889.

[8] Hansen, M. and Yu, B. (2001). Model selection and Minimum Description Length principle. *J. Amer. Statist. Assoc.* , **96**, 746-774.

[9] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, **90**, 773–795.

[10] O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics: Bayesian Inference. Vol 2B.* New York: John Wiley & Sons.

[11] Peterson, J. J. (1986). A note on some model selection criteria. *Stat. and Prob. Letters*, **4**, 227–230.

[12] Rissanen, J. (1989). *Stochastic complexity and statistical inquiry.* Singapore: World Scientific.

[13] Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory*, **42**, 48–54.

[14] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 4611–464.

[15] Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

[16] T. Speed and B. Yu (1993). Model selection and prediction: normal regression. *J. Inst. Statist. Math.* **45**, 35-54.

[17] West, M. (1985). Generalized linear models: scale parameters, outlier accomodation and prior distributions. *Bayesian Statistics 2*, pp. 531-558, Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (Eds). Elsvier Science Publishers B. V. (North-Holland).

[18] Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Eds. P.K. Goel and A. Zellner, 233–243. Amsterdam: North-Holland.