Minimum Description Length Model Selection Criteria for Generalized Linear Models

Mark H. Hansen and Bin Yu

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. In the latter case, we apply Efron's [6] double exponential family characterization of GLMs. Standard Laplace approximations are then employed to derive computationally tractable selection rules. Each constructed criterion has 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" the selection "extremes" 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.

Keywords: 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. As a principle, maximum likelihood is not well suited for this problem as it suggests choosing the largest model under consideration. Following this strategy, we tend to overfit the data and choose models that have poor predictive power. Model selection emerged as a field in the 1970s, introducing procedures that "corrected" the maximum likelihood approach. The most famous and widely used criteria are *An Information Criterion* (AIC) of Akaike [1, 2] and the *Bayesian Information Criterion* (BIC) of Schwarz [15]. They both take the form of a penalized maximized likelihood, but with different penalties: AIC adds 1 for each additional variable included in a model, while BIC adds $\log n/2$, where n is the sample size. Theoretical and simulation studies (*cf.* Shibata [16], Speed and Yu [18], and references therein), mostly in the regression case, have revealed that when the underlying model is finite-dimensional (specified by a finite number of parameters), BIC is preferred; but when it is infinite-dimensional, AIC performs best. Unfortunately, in practical applications we rarely have this level of

information about how the data were generated, and it is desirable to have selection criteria which perform well independent of the form of the underlying model. That is, we seek criteria which adapt automatically to the situation at hand. In this paper, we derive such adaptive model selection criteria for generalized linear models (GLMs) under the Minimum Description Length (MDL) framework. With MDL we find several generic prescriptions or "forms" for constructing such selection criteria. In this paper, we focus on one MDL form that is based on mixtures.

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 [13], Barron *et al.* [3], and Hansen and Yu [8]) discriminates between competing model classes based on the *complexity* of each description. Precisely, the Minimum Description Length (MDL) Principle recommends that we

Choose the model that gives the shortest description of data.

While there are many kinds of descriptions and many ways to evaluate their complexity, we follow Rissanen [13] and use a *code length* formulation based on the candidate model.

To make this more precise, we first recall that for each probability distribution Q on a finite set $\mathcal{A}$ there is an associated *code* that prepares elements of $\mathcal{A}$ for transmission across some (noiseless) communication channel. We consider binary codes, meaning that each codeword is a string of 0's and 1's. It is possible to find a code so that the number of bits (the number of 0's and 1's in a codeword) used to encode each symbol of $a \in \mathcal{A}$ is essentially $-\log_2 Q(a)$; that is, $-\log_2 Q$ can be thought of as a *code length function*. Huffman's algorithm [5] takes a distribution Q and produces a so-called *prefix code* with the right length function.[1] Conversely, any integer-valued function L corresponds to the code length of some binary prefix code if and only if it satisfies Kraft's inequality

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

see Cover and Thomas [5] for a proof. Therefore, given a prefix code on $\mathcal{A}$ with length function L, we can define a distribution on $\mathcal{A}$ as follows:

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

With Kraft's inequality, we find a correspondence between codes and probability distributions. In what follows, we work with natural logs and take $-\log Q$ to be an idealized code length.

---

[1] While the details are beyond the scope of this short paper, the interested reader is referred to Hansen and Yu [8] and Cover and Thomas [5].

One of the early problems in information theory involves transmitting symbols that are randomly generated from a probability distribution P defined on $\mathcal{A}$. Let $A \in \mathcal{A}$ denote a random variable with this distribution. Now, from the discussion in the previous paragraph, any code defined on $\mathcal{A}$ can be associated with an idealized code length function $-\log Q$ for some distribution Q. With this setup, the expected code length for symbols generated from P is given by $-E \log Q(A) = -\sum_a P(a) \log Q(a)$. By Jensen's inequality, we see that the shortest code length is achieved by a code that has $-\log P$ as its idealized length function. That is, the expected code length is bounded from below by $-E \log P(a) = -\sum_a P(a) \log P(a)$, the entropy of P. In the literature on information theory, this fact is known as Shannon's Inequality.

In this paper, we focus on descriptions of data that consist of probability models, and compare them based on the efficiency of the corresponding code in terms of improvements in code length relative to the entropy of the data generating process. When the competing models are members of a parametric family, using MDL to select a model, or rather, to estimate a parameter, is equivalent to maximum likelihood estimation (when the cost of transmitting the parameter estimate is fixed). To compare different model classes, different parametric families, or carry out model selection from among several candidate model classes, efficient codes for each class need to *fairly represent* its members. We do not elaborate on this idea, but instead comment that it is possible to demonstrate rigorously that several coding schemes achieve this fairness and hence provide valid selection criteria (for, say, i.i.d. or time series observations). We refer readers to Barron *et al.* [3] and Hansen and Yu [8].

Among the schemes that yield valid selection criteria, the best known is the so-called *two-stage code*, in which we first encode the maximum likelihood estimate (MLE) of the parameters in the model, and then use the model with the MLE to encode the data (say, via Huffman's algorithm described above). Hence this form is a penalized likelihood, and to first order is exactly the same as BIC. Other forms of MDL include predictive, mixture and normalized maximum likelihood (NML). The predictive form makes the most sense when the data come in sequentially and has a close connection to prequential inference; the mixture codes are described in more detail in the next section; the NML form is new and evolving, and code length expressions are known only in a few special cases, including the binomial model and Gaussian linear regression (*cf.* Rissanen [14], Barron *et al.* [3], and Hansen and Yu [8]).

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 (so named because of its use of Zellner's g-prior [20]) from Hansen and Yu [8] 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 [7] in the context of a Bayesian analysis. Section3.3 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 a version of the mixture form gMDL for GLMs. In this case,

normal or normal-inverse gamma distributions are used to construct a mixture model, 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 [11] in the context of a Bayesian analysis. To account for dispersion effects, we use Efron's [6] double exponential family characterization of GLMs as the likelihood. Standard Laplace approximations are employed to derive computationally tractable selection rules. Each constructed criterion has 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.

2   Mixture MDL

In this paper, we 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 (or conditional densities) for Y, indexed by a 0-1 binary vector $\gamma = (\gamma_1, \ldots, \gamma_K)$; each model depends on a subset of the covariates corresponding to 1's in the model index vector γ. Generically, we let $\mathcal{M}_\gamma$ denote a simple model class with dimension $k_\gamma = \sum_{j=1}^{K} \gamma_j$, which depends on the predictors $(X_1, \ldots, X_K)$ through the linear combination

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

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

For simplicity, we now drop the subscript γ except in places where a reminder seems necessary. The reader should interpret the model class $\mathcal{M}$, its dimension k, the design matrix X, and the parameters θ and β 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 $\mathcal{M}$

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

where w is a probability density function on θ. This integral has a closed form expression when $f_\theta(\cdot|X)$ 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 $y \in \mathcal{Y}^n$. In this setting, we assume that both sender and receiver know about the covariates X, and we only have to transmit y. As an example, suppose $\mathcal{Y} = \{0, 1\}$ so that 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 $y \in \{0, 1\}^n$. From the discussion in Section 1, we can apply Huffman's algorithm to build a code that has the (idealized) length function $L(y) = -\log_2 m(y|X)$ for all $y \in \mathcal{Y}^n$. This means that the number of bits required to transmit any $y \in \{0, 1\}^n$ is essentially $-\log_2 m(y|X)$. The MDL principle then distinguishes between candidate model classes based on the associated length function $L(Y)$, the number of bits required to transmit the observed data Y. As mentioned earlier, we have chosen to use base e in the log for our derivations.

In Section 1, we only considered building codes for finite sets of symbols. When $Y_i \in \mathcal{Y} \subset \mathbb{R}$, $i = 1, \ldots, n$, is a continuous response, we form an approximate length function by first discretizing the set $\mathcal{Y}$. That is, given a precision δ we obtain the description length

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

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

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

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 [8].

When choosing between two model classes, the mixture form of MDL (with fixed hyperparameters) is equivalent to a Bayes factor (Kass and Raftery [9]) based on the same distributions on the parameters spaces. As we 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 Y from $\mathcal{M}$. This distinction also allows more freedom in choosing w, and has spawned a number of novel applications in engineering.

3   Regression

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

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

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 $(X_1, \dots, X_K)$, we restored the γ notation in the above equation. For simplicity, however, from this point on, we 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 X_K, but instead we entertain model classes because they are capable of capturing the major features observed in the observed data string 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 θ in the mixture (3) consists only of the coefficients β; while in the latter, θ involves both β and σ^2.

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

3.1   Known error variance σ^2

Here, we take $\theta = \beta$ and let $w(\beta|X)$ 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

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

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 $V = c(X^tX)^{-1}$ so that

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

where $FSS = y^tX(X^tX)^{-1}X^ty$ is the usual fitted sum of squares corresponding to the OLS estimate $\widehat{\beta} = (X^tX)^{-1}X^tY$. This particular choice of distribution is often attributed

to Zellner [20] 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 β is $\frac{c}{1+c}\widehat{\beta}$.

In (8) we have highlighted the dependence of the mixture on the scaling parameter *c*. George and Foster [7] studied various approaches to setting *c*, establishing that certain values lead to well-known selection criteria like AIC and BIC.[2] Ultimately, they propose an empirical Bayes approach, selecting an estimate $\hat{c}$ via maximum likelihood. Hansen and Yu [8] 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 [8] 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 adds to the code length, a charge that we make explicit by writing

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

Following Rissanen [13], the cost $L(c)$ is taken to be $\frac{1}{2}\log n$.[3] Minimizing (9) with respect to *c* gives

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

and substituting into (8) yields a code length (9) of the form

$$L(y) = \begin{cases} \frac{y'y - FSS}{2\sigma^2} + \frac{k}{2}\left[1 + \log\left(\frac{FSS}{\sigma^2 k}\right)\right] + \frac{1}{2}\log n & \text{for } FSS > k\sigma^2 \\ \\ \frac{y'y}{2\sigma^2} & \text{otherwise.} \end{cases} \tag{10}$$

When the minimizing value of *c* is zero, the prior on β 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 criterion of George and Foster [7] is basically the same, but leaves off this extra term.

---

[2]A similar calibration between Bayesian methods and well-known selection criteria can also be found in Smith and Spielgelhalter [17].

[3]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. George and Foster [7] advocate estimating σ^2 then applying the form (10); however, we prefer to assign a distribution to σ^2 and incorporate it into the mixture. Following Hansen and Yu [8], we employ a conjugate normal-inverse gamma distribution to form the mixture code; that is, $1/\sigma^2$ has a gamma distribution with shape parameter a; and given σ^2, β is normal with mean zero and variance $\sigma^2 V$. Setting $\tau = \sigma^2$, these densities are given by

$$w(\beta, \tau) \propto \tau^{-\frac{(k+3)}{2}} \exp\left[\frac{-\beta^t V^{-1}\beta + a}{2\tau}\right], \tag{11}$$

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

$$
\begin{aligned}
-\log m(y|X, a, V) &= \frac{1}{2}\log|V^{-1} + X^t X| + \frac{1}{2}\log|V| - \frac{1}{2}\log a \\
&\quad + \frac{n+1}{2}\log\left(a + y^t y - y^t X\left(V^{-1} + X^t X\right)^{-1}X^t y\right)
\end{aligned}
\tag{12}
$$

where we have ignored terms that do not depend on our particular choice of model. The derivation of $m(y|X, a, V)$, the marginal or predictive distribution of y, is standard and can be found in O'Hagan [10].

Our approach to handling the hyperparameter a is the same as that in the previous section. Minimizing (12) with respect to a we find that $\hat{a} = (y^t y - y^t X(V^{-1} + X^t X)^{-1}X^t y)/n$ which leaves

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

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

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

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

$$\hat{c} = \max\left(F - 1, 0\right) \quad \text{with} \quad F = \frac{FSS}{kS}, \tag{15}$$

where F is the usual F-ratio for testing the hypothesis that each element of β is zero, and $S = RSS/(n-k)$. The truncation at zero in (15) rules out negative values of the prior variance. Rewriting (15), 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 β becomes a point mass at zero, effectively producing the "null" mixture model[4] corresponding to zero regression effects. Substituting the optimal value of $\hat{c}$ into (14), we arrive at a final mixture form

$$\text{gMDL} = \begin{cases} \frac{n}{2}\log S + \frac{k}{2}\log F + \log n, & R^2 \geqslant k/n \\ \frac{n}{2}\log\left(\frac{y'y}{n}\right) + \frac{1}{2}\log n, & \text{otherwise}. \end{cases} \tag{16}$$

Note that we have added the cost to code the hyperparameters a and c, producing an extra $\log n$ and $(1/2)\log n$ in the upper and lower expressions, respectively.

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

$$\text{AIC} = \frac{n}{2}\log RSS + k \quad \text{and} \quad \text{BIC} = \frac{n}{2}\log RSS + \frac{k}{2}\log n. \tag{17}$$

Comparing these with (16), we see that the essential difference is in the penalty. Both AIC and BIC have data independent 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 the figure of merit to be prediction error, then AIC has been shown both through theory and simulation studies to be optimal in this setting. When the true, data generating mechanism is finite-dimensional (and is included among the candidates being compared), the stronger penalty of BIC tends to perform better. For this kind of problem, selection criteria may also be judged 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. BIC has been shown to perform optimally in this setting.

We 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 [4], we assume that $X_i = (X_{i1}, X_{i2}, \ldots, X_{ij}, \ldots)$ are Gaussian, zero-mean random vectors and let

$$\sigma_k^2 = \text{var}\left(\sum_{j=k+1}^{\infty} \beta_j X_{ij} \,\middle|\, X_{i1}, \ldots, X_{ik}\right).$$

---

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

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 [4] and the fact that $\frac{1}{n}||y||^2 = (\sigma^2 + \sigma_0^2)(1 + o_p(1))$, it is straightforward to establish the following two results.

Theorem 9
The quantity F in (15) satisfies $F = [\frac{n}{k}\frac{\sigma_0^2 - \sigma_k^2}{\sigma_k^2 + \sigma^2} + 1](1 + o_p(1))$ where $o_p(1) \to 0$ in probability uniformly over $0 \leqslant k \leqslant n/2$.

Corollary 3
If the model is finite-dimensional and the maximum dimension of the models examined $K = K_n = o(n)$, then gMDL is consistent and is also prediction-optimal.

The above theorem presents an expansion of the data dependent-penalty of gMDL, and the corollary establishes that gMDL enjoys the same optimality as BIC when the model is finite-dimensional. When $\sigma_k^2 > 0$ for all k, the underlying model is infinite-dimensional. In this case, the quantity F/n can be viewed as the average signal to noise ratio for the fitted model. Adjusting the penalty with $(k/2)\log F/n$, gMDL is able to adapt to perform well in terms of prediction in both domains, finite- or infinite-dimensional. 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 show that the newly derived MDL-based criteria for GLMs are also adaptive.

4   Generalized Linear Models

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

$$f(y) = \exp\left(\frac{y\psi - b_1(\psi)}{b_2(\phi)} + b_3(y,\phi)\right), \quad y \in \mathcal{Y}, \tag{18}$$

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 (18) 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 (18),

$$\begin{aligned} E(Y) &= \mu &= b_1'(\psi) \\ \mathrm{var}(Y) &= \sigma^2 &= b_1''(\psi)b_3(\phi). \end{aligned} \tag{19}$$

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 $(X_1, \dots, X_K)$. Each model class

corresponds to some value of the binary vector $\gamma = (\gamma_1, \ldots, \gamma_K)$, and we relate the mean μ of Y to a subset of the covariates via the linear predictor

$$\eta = g(\mu), \qquad \text{for } \eta = \sum_{j:\gamma_j=1} \beta_j X_j, \tag{20}$$

where g is a one-to-one, continuously differentiable transformation known as the link function. Using (19) and (20) we see that $\eta = g(b'(\psi))$.[5] Again we let $\beta_\gamma = (\beta_j)_{j:\gamma_j=1}$ denote the vector of regression coefficients and $k_\gamma = \sum \gamma_j$ its dimension. The unknown parameters associated with this model are denoted θ_γ and include both β_γ as well as a possible dispersion effect ϕ. We observe data of the form (Y_i, X_i) for $i = 1, \ldots, n$ where $X_i = (X_{i1}, \ldots, X_{iK})$ and again X_K is the $n \times K$ full design matrix $[X_K]_{ij} = X_{ij}$. We let X_γ refer to a submatrix of X_K consisting of only those columns j for which $\gamma_j = 1$. Let $f_{\theta_\gamma}(y|X_\gamma)$ denote the density for Y based on model class γ.

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

$$m(y|X) \approx (2\pi)^{\frac{k}{2}} |-H^{-1}(\tilde{\beta})|^{\frac{1}{2}} f_{\tilde{\beta}}(y|X) w(\tilde{\beta}), \tag{21}$$

where H is the Hessian of $h(\beta) = \log f_\beta(y|X) + \log w(\beta)$ and $\tilde{\beta}$ is the posterior mode of β. In working with this form, we repeatedly make use of the Fisher information matrix $I(\beta) = X^t W(\beta) X$, where W is a diagonal weight matrix. Note that for GLMs, the observed Fisher information is the same as the Fisher information when we use the canonical parameterization.

Form (21) is still difficult to work with in practice because there is typically no closed-form expression for the posterior mode. We 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 derive a criterion that first appeared in Peterson [11]. 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 follows closely the case of σ^2 known for regression. Let β be the MLE of β, and assume that the prior $w(\beta)$ is normal with mean zero and variance-covariance V. Then, we can approximate $\tilde{\beta}$ via a single Newton step

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

---

[5]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 $H(\beta) = -(I(\beta) + V^{-1})$, where $I(\widehat{\beta})$ is the Fisher information evaluated at $\widehat{\beta}$. We now focus on the case where the prior variance-covariance matrix for β is simply $cI(\widehat{\beta})^{-1}$. 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 $\widehat{\beta}$. From a strict coding perspective this is hard to accept; it would imply that sender and receiver both know the coefficient $\widehat{\beta}$ (or at least $I(\widehat{\beta})$). Nonetheless, it is instructive to follow this line of analysis and compare it with the results of the previous section. For $V = cI(\widehat{\beta})^{-1}$ we find that the one-step Newton-Raphson iteration gives

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

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

Continuing with the expression (21), we find that

$$
\begin{aligned}
\log w(\tilde{\beta}) &\approx \log w\left(\frac{c}{1+c}\widehat{\beta}\right) \\
&= -\frac{k}{2}\log 2\pi + \log|cI(\widehat{\beta})| - \frac{1}{2}\frac{c}{(1+c)^2}\widehat{\beta}^t I(\widehat{\beta})\widehat{\beta}
\end{aligned}
\tag{22}
$$

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

$$
\begin{aligned}
\log f_{\widehat{\beta}}(y|X) &\approx \log f_{\widehat{\beta}}(y|X) - \frac{1}{2}(\tilde{\beta}-\widehat{\beta})^t I(\widehat{\beta})(\tilde{\beta}-\widehat{\beta}) \\
&= \log f_{\widehat{\beta}}(y|X) - \frac{1}{2}\frac{1}{(1+c)^2}\widehat{\beta}^t I(\widehat{\beta})\widehat{\beta}.
\end{aligned}
\tag{23}
$$

Combining (22) and (23) we arrive at the expression

$$
\begin{aligned}
\log f_{\widehat{\beta}}(y|X) + \log w(\tilde{\beta}) &\approx \log f_{\widehat{\beta}}(y|X) - \frac{1}{2}\frac{1}{1+c}\widehat{\beta}^t I(\widehat{\beta})\widehat{\beta} \\
&\quad -\frac{k}{2}\log 2\pi + \frac{k}{2}\log c + \frac{1}{2}\log|I(\widehat{\beta})|.
\end{aligned}
\tag{24}
$$

Finally, collecting terms in (21) we find an expression for the code length given c

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

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

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

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

$$
L(y) = \begin{cases} -\log f_{\widehat{\beta}}(y|X,\widehat{\beta}) + \frac{k}{2}\left[1+\log\left(\frac{\widehat{\beta}' I(\widehat{\beta})\widehat{\beta}}{k}\right)\right] + \frac{1}{2}\log n & \text{for } \widehat{\beta}' I(\widehat{\beta})\widehat{\beta} > k \\[2ex] -\log f_0(y|X) & \text{otherwise.} \end{cases}
$$

The function $f_0(y|X)$ represents the log-likelihood when all 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 value, forcing consideration of over-dispersed models. There are several ways to introduce extra variability into the form (18), many of which are primarily meant as computational devices. Efron [6] 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 [19].

To understand this form, we have to first rewrite the log-likelihood for a GLM in terms of its mean vector $l(y|\mu)$, where $\mu = (\mu_1,\ldots,\mu_n)$. Now, using this notation, without the restriction (20) on the mean, the maximum value of the log-likelihood is simply $l(y|y)$. We then define the deviance as the difference

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

where β is the vector of regression coefficients that yield μ through (20). To incorporate a dispersion parameter, Efron [6] motivates the use of

$$
\tau^{-n/2} e^{l(y|\mu)/\tau + (1-1/\tau)l(y|y)} \tag{25}
$$

as an (approximate) likelihood. Technically, this expression should include a normalizing constant $C(\tau,\beta)$. Following Efron [6], however, it can be shown that $C(\tau,\beta) = 1 + O(n^{-1})$, and hence can be ignored for reasonable sample sizes. Rewriting (25), we work with

$$
\tau^{-n/2} e^{-\frac{(2l(y|y)-2l(y|\mu))}{2\tau}} e^{l(y|y)} = \tau^{-n/2} e^{-\frac{D(y|\beta)}{2\tau}} e^{l(y|y)}. \tag{26}
$$

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 β, τ and y is given by

$$
\tau^{-1-(n+1)/2}\frac{e^{-a/2\tau}}{\sqrt{\pi}}\sqrt{\frac{a}{2}}(2\pi\tau)^{-k/2}|V^{-1}|^{1/2}e^{\frac{-\beta' V^{-1}\beta - D(y|\beta)}{2\tau}}. \tag{27}
$$

To integrate out β, we use the Laplace method again which this time yields

$$\frac{\tau^{-1-(n+1)/2}}{\sqrt{\pi}}\sqrt{\frac{a}{2}}|V^{-1}|^{1/2}|V^{-1}+I(\widehat{\beta})|^{-1/2}e^{\frac{-a-\tilde{\beta}^tV^{-1}\tilde{\beta}-D(y|\tilde{\beta})}{2\tau}}, \qquad (28)$$

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

$$-\log m(y|a,V) = \frac{n+1}{2}\log\left(a+\tilde{\beta}^tV^{-1}\tilde{\beta}+D(y|\tilde{\beta})\right)$$
$$-\frac{1}{2}\log a+\frac{1}{2}\log|V|+\frac{1}{2}\log|V^{-1}+I(\tilde{\beta})|. \qquad (29)$$

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

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

We have now obtained a usable criterion for model selection. Specifying V, we can compute $\tilde{\beta}$ with simple Newton-Raphson iterations. In the regression analysis, we used Zellner's g-prior for β which led to a closed-form selection criterion. The analog in this case is $V = cI^{-1}(\widehat{\beta})$. For a GLM, this choice is somewhat unsettling because $I(\widehat{\beta})$ 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 $\widehat{\beta}$, or at least $I(\widehat{\beta})$. Recall that for a GLM, the Fisher information matrix takes the form $X^tW(\widehat{\beta})X$ where W is a diagonal weight matrix. One simple alternative is to take $V = c(X^tX)^{-1}$, or $V = c1$, where 1 is the identity matrix. In each of these cases, we must either approximate the $\tilde{\beta}$ or iterate to find it. We consider both kinds of selection criteria.

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

$$\tilde{\beta} \approx \frac{c}{1+c}\widehat{\beta} \qquad (30)$$

and

$$\frac{k}{2}\log(1+c) + \frac{n}{2}\log\left(\frac{1}{1+c}\widehat{\beta}^tI(\widehat{\beta})\widehat{\beta}+D(y|\widehat{\beta})\right). \qquad (31)$$

Here we have substituted in the one-step Newton-Raphson approximation for $\tilde{\beta}$ and have approximated the deviance $D(y|\tilde{\beta})$ by a Taylor expansion around $\widehat{\beta}$ and used a relation from Raftery [12]. Maximizing with respect to c yields

$$\hat{c} = \max\left(F - 1, 0\right) \qquad (32)$$

where

$$F = \frac{(n-k)\widehat{\beta}^t I(\widehat{\beta})\widehat{\beta}}{kD(y|\widehat{\beta})}.$$

This then gives the form

$$L(y) = \begin{cases} \frac{n}{2}\log\frac{D(y|\widehat{\beta})}{n-k} + \frac{k}{2}\log F + \log n & \text{if } F > 1 \\ \frac{n}{2}\log\frac{D(y|0)}{n} + \frac{1}{2}\log n & \text{otherwise,} \end{cases}$$

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

For the other choices of $V^{-1} = c\Sigma$, there is not a closed-form expression for the maximizing c. Instead, we can perform a search, but this is best done in conjunction with finding $\tilde{\beta}$. It is also possible to use the approximate $\tilde{\beta}$ (30) to derive a simple iteration to find c. In this case, we find

$$c = \frac{kR_c}{n\widehat{\beta}^t I(\widehat{\beta})\left(I(\widehat{\beta})+c\Sigma\right)^{-1}\Sigma\left(I(\widehat{\beta})+c\Sigma\right)^{-1}I(\widehat{\beta})\widehat{\beta} + R_c\text{trace}\left(c1 + I(\widehat{\beta})\Sigma^{-1}\right)} \tag{33}$$

where

$$R_c = D(y|\widehat{\beta}) + c\widehat{\beta}^t\Sigma\widehat{\beta} - c^2\widehat{\beta}^t\Sigma(I(\widehat{\beta})+c\Sigma)^{-1}\Sigma\widehat{\beta}. \tag{34}$$

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 X. In the first, each column consists of $n = 100$ 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

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

Here, we took $\rho = 0.75$. Then Y was generated by 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. Table 1 gives 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 $V^{-1} = c^{-1}I(\widehat{\beta})$ (Section 4.2); Column 5 corresponds to mixture MDL with a fixed dispersion parameter ϕ and hence a normal mixing distribution again with $V^{-1} = c^{-1}I(\widehat{\beta})$ (Section 4.1); Columns 6 and

Table 1: Classification errors for the different selection criteria.

Coefficients	ρ	Bayes Rate	Mix ϕ	$\phi = 1$	BIC	AIC	1 Iter	X^tX Iter	1 search	X^tX 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

7 are BIC and AIC (17). Columns 8 through 11 also make use of the normal-inverse gamma distribution but with different choices of the variance-covariance matrix V^{-1}; $c^{-1}1$ for 8 and 10, and $c^{-1}X^tX$ for 9 and 11. Columns 8 and 10 differ only in how we estimate $\tilde{\beta}$ and $\hat{c}$; in the first case the iteration (33) is used, while in the second a full search is performed to identify both $\tilde{\beta}$ and the appropriate value of c. The same holds for Columns 9 and 11, but with the different variance-covariance matrix.

Table 1 shows that most of the selection criteria behave the same, at least in terms of classification error; this 0-1 error is very robust. In Table 2 we illustrate the types of models selected by each scheme. The first column identifies the simulations from Table 1. The second column presents a model summary 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 the first panel of Table 2, the true model $(2, 0, 0, 0, 0)$ consists of only one effect. The heading "1-0" represents the correct model and is marked with a "*", while the column "1-1" means that one extra term was included. From this table, we see that the three MDL criteria (Columns 9, 11 and 12) adapt to either AIC or BIC depending on which performs better in all 8 set-ups. Column 10 seemed to have some problems, and we believe this is because the iterations (33) failed to converge properly (possibly due to the approximations used to generate the form). Finally, we see that the columns using $I(\hat{\beta})$ can perform poorly (those denoted Mixture ϕ and $\phi = 1$). Recall that we derived these forms even though their reliance on $\tilde{\beta}$ violates the basic coding ideas behind MDL.

We consider the cases in more depth, starting with the first panel of Table 2. Here "truth" is a small model, $(2, 0, 0, 0, 0)$, an ideal case for BIC. Clearly, BIC selects the right model more often than the other procedures. The mixture MDL procedures that use variance-covariance matrices other than $I(\hat{\beta})$ also perform quite well. In terms of

Table 2: Summarizing the number of times different sized models were selected for a sample of simulation runs given in Table 1.

Coefficients	Model Summary	Mix ϕ	$\phi = 1$	BIC	AIC	1 Iter	X^tX Iter	1 search	X^tX search
$\beta = (2,0,0,0,0)$	0-1	0	0	0	0	0	0	0	0
	0-2	0	0	0	0	0	0	0	0
	0-3	0	0	0	0	0	0	0	0
	0-4	0	0	0	0	0	0	0	0
*	1-0	131	134	215	121	176	179	183	179
	1-1	70	72	31	85	55	53	51	51
	1-2	37	37	4	40	17	18	15	19
	1-3	12	7	0	4	2	0	1	1
	1-4	0	0	0	0	0	0	0	0
$\beta = (3,2,2,0,0)$	0-1	0	0	0	0	0	0	0	0
	0-2	0	0	0	0	0	0	0	0
	1-0	0	0	0	0	0	0	0	0
	1-1	0	0	0	0	0	0	0	0
	1-2	0	0	0	0	0	0	0	0
	2-0	0	0	0	0	0	3	0	0
	2-1	0	0	1	0	0	0	0	0
	2-2	0	0	0	0	0	0	0	0
*	3-0	111	134	227	173	176	71	184	180
	3-1	103	93	20	71	49	31	61	64
	3-2	36	23	2	6	25	145	5	6
$\beta = (5,1,1,1,1)$	1-0	0	0	4	0	0	1	0	0
	2-0	0	1	35	2	4	19	3	3
	3-0	9	12	64	24	33	13	23	23
	4-0	56	70	89	94	79	18	86	89
*	5-0	185	167	58	130	134	199	138	135

test error, each procedure is about the same. Overall, we can recommend the MDL-based criteria in terms of their ability to adapt and select concise models.

In the second panel of Table 2, the coefficient vector is $(3,2,2,0,0)$, a middle-ground case. The $I(\widehat{\beta})$ criteria perform rather poorly, as does the X^tX case with iterations (33) to find $\hat{c}$. In the latter case, the poor performance is even reflected in the prediction error. We intend to examine whether the approximation that led to (33) caused the problem, or if it was poor starting values for the iterations.

Finally, in the last panel of Table 2, we consider a "full" 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.

Acknowledgments

Dedication

This paper is dedicated to Terry Speed, who introduced B. Yu to MDL 15 years ago. By encouraging the highest academic standards, Terry is an inspiration to all of Berkeley's students.

Mark H. Hansen, Statistics Research, Bell Laboratories, Lucent Technologies,
cocteau@bell-labs.com

Bin Yu, Department of Statistics, University of California, Berkeley,
binyu@stat.berkeley.edu

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] H. Akaike. An objective use of Bayesian models. *Annals of the Institute of Statistical Mathematics*, 29:9–20, 1977.

[3] A. Barron, J. Rissanen, and B. Yu. Minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory. (Special Commemorative Issue: Information Theory: 1948-1998)*, 44:2743–2760, 1998.

[4] L. Breiman and D. Freedman. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78:131–136, 1983.

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

[6] B. Efron. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81:709–721, 1986.

[7] E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87:731–747, 2000.

[8] M. Hansen and B. Yu. Model selection and minimum description length principle. *Journal of the American Statistical Association*, 96:746–774, 2001.

[9] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

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

[11] J. J. Peterson. A note on some model selection criteria. *Statistics and Probability Letters*, 4:227–230, 1986.

[12] A. E. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83:251–266, 1996.

[13] J. Rissanen. *Stochastic complexity and statistical inquiry*. World Scientific, Singapore, 1989.

[14] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42:48–54, 1996.

[15] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:4611–464, 1978.

[16] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

[17] A. F. M. Smith and D. J. Spielgelhalter. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B*, 42:213–220, 1980.

[18] T. Speed and B. Yu. Model selection and prediction: normal regression. *Journal of the Institute of Statistical Mathematics*, 45:35–54, 1993.

[19] M. West. Generalized linear models: scale parameters, outlier accomodation and prior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2*, pages 531–558, Amsterdam, 1985. North-Holland.

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