Minimum Description Length Methods in Bayesian Model Selection: Some Applications

Mohan Delampady
Statistics and Mathematics Unit, Indian Statistical Institute, Bangalore, India
Email: mohan.delampady@gmail.com

Received January 8, 2013; revised February 10, 2013; accepted February 26, 2013

Copyright © 2013 Mohan Delampady. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT
Computations involved in Bayesian approach to practical model selection problems are usually very difficult. Computational simplifications are sometimes possible, but are not generally applicable. There is a large literature available on a methodology based on information theory called Minimum Description Length (MDL). It is described here how many of these techniques are either directly Bayesian in nature, or are very good objective approximations to Bayesian solutions. First, connections between the Bayesian approach and MDL are theoretically explored; thereafter a few illustrations are provided to describe how MDL can give useful computational simplifications.

Keywords: Bayesian Analysis; Model Selection; Minimum Description Length; Hierarchical Bayes; Bayesian Computations

1. Introduction
Bayesian computations can be difficult, in particular those in model selection problems. For instance, it may be noted that learning the structure of Bayesian networks is in general of the computational complexity type NP-complete ([1,2]). It is therefore meaningful to consider alternative computationally simpler solutions which are approximations to Bayesian solutions. Sometimes direct computational simplifications are possible, as shown, for example, in [3], but often approaches arising out of different methodologies are needed. We discuss some aspects of Minimum Description Length (MDL) methods with this point of view. Another important reason for exploring these methods is that there is a substantial literature on this topic available in engineering and computer science with potential applications in statistics. We will not, however, explore certain other aspects of MDL such as the “Normalized Maximum Likelihood (NML)” introduced by [4] which do not seem to be in the spirit of the Bayesian approach that we have taken here.

The discussion below is organized as follows. In Section 2 we briefly describe the MDL principle and then indicate in Sections 3 and 4 how it applies to model fitting and model checking. It is shown that a particular version of MDL is equivalent to the Bayes factor criterion of model selection. Since this is computationally difficult most often, some approximations are desirable, and it is next shown how a different version of MDL can provide such an approximation. Following this discussion, new applications are presented in Section 5. Specifically, MDL approach to step-wise regression in Section 5.1, wavelet thresholding in 5.2 and a change-point problem in 5.3 are described.

2. Minimum Description Length Principle
The MDL approach to model fitting can be described as follows (see [5,6]). Suppose we have some data. Consider a collection of probability models for this set of data. A model provides a better fit if it can provide a more compact description for the data. In terms of coding, this means that according to MDL, the best model is the one which provides the shortest description length for the given data. The MDL approach as discussed here is also related to the Minimum Message Length (MML) approach of [7]. See [8,9] for connections to information theory and other related details.

If data \( x \) is known to arise from a probability density \( p \), then (see [10] or [11]) the optimal code length (in an average sense) is given by \( -\log p(x) \). (Here log is logarithm to the base 2.) This is the link between description length and model fitting.

The optimal code length of \( -\log p(x) \) is valid only in the discrete case. To handle the continuous case later, discretize \( x \) and denote it by \( [x]=x_\delta \) where \( \delta \)
denotes the precision. In effect we will then be considering
\[ P\left[ \frac{1}{2} x \leq X \leq \frac{1}{2} x \right] = \int_{\frac{1}{2} x}^{\frac{1}{2} x} p(u)du \approx \delta p(x) \]
instead of \( p(x) \) itself as far as coding of \( x \) is considered when \( x \) is one-dimensional. In the \( r \)-dimensional case, we will replace the density \( p(x) \) by the probability of the \( r \)-dimensional cube of side \( \delta \) containing \( x \), namely \( p\left[ \left[ x \right] \delta \right] \approx p(x)\delta^r \), so that the optimal code length changes to \( -\log p(x) - r \log \delta \).

3. MDL for Estimation or Model Fitting

Consider data \( x = x^0 = (x_1, x_2, \cdots, x_n) \), and suppose
\[ \mathcal{F} = \left\{ f(x^0 | \theta) : \theta \in \Theta \right\} \]
is the collection of models of interest. Further, let \( \pi(\theta) \) be a prior density for \( \theta \). Given a value of \( \theta \) (or a model), the optimal code length for describing \( x \) is \( -\log f(x^0 | \theta) \), but since \( \theta \) is unknown, its description requires a further \( -\log \pi(\theta) \) bits on average. Therefore the optimal code length is obtained upon minimizing
\[ DL(\theta) = -\log \pi(\theta) - \log f(x^0 | \theta), \tag{1} \]
so that MDL amounts to seeking that model which minimizes the sum of
- the length, in bits, of the description of the model, and
- the length, in bits, of data when encoded with the help of the model.

Now note that the posterior density of \( \theta \) given the data \( x^0 \) is
\[ \pi(\theta | x^0) = \frac{f(x^0 | \theta) \pi(\theta)}{m(x^0)}, \tag{2} \]
where \( m(y) \) is the marginal or predictive density. Therefore, minimizing
\[ DL(\theta) = -\log \pi(\theta) - \log f(x^0 | \theta) \]
over \( \theta \) is equivalent to maximizing \( \pi(\theta | x^0) \). Thus MDL for estimation or model fitting is equivalent to finding the highest posterior density (HPD) estimate of \( \theta \). Note, however, that a prior \( \pi \) is needed for these calculations. The approach that a Bayesian adopts in specifying the prior is not, in general, what is accepted by practitioners of the MDL approach. Therefore, the equivalence of MDL and HPD approaches is either subject to accepting the same prior, or as an asymptotic or similar approximation. MDL mostly prefers an approximately uniform prior when \( \Theta \subset \mathbb{R}^k \) for some fixed \( k \) (same \( k \) across all models), leading to the maximum likelihood estimate (MLE). The case of \( \mathcal{F} \) having model parameters of different dimensions is different and is interesting. This can be easily seen in the continuous case upon discretization. Now denote \( DL \) by \( DL^r \) and \( \theta \) by \( \theta^k = (\theta_1, \theta_2, \cdots, \theta_k) \). Then
\[ DL^r(\theta^k) = -\log \pi\left[ \left[ \theta^k_{-x} \right]_{\delta_x} \right] - \log f\left( \left[ x^0_{-x} \right]_{\delta_x} \right) \]
\[ = -\log \pi\left( \left[ \theta^k_{-x} \right]_{\delta_x} \right) - k \log \delta_x - \log f\left( \left[ x^0_{-x} \right]_{\delta_x} \right) \]
\[ = -n \log \delta_x \]
\[ \approx -\log \left( \pi(\theta^k) \right) - k \log \delta_x - \log f\left( x^0 | \theta^k \right) - n \log \delta_f. \]

Here \( \delta_f \) and \( \delta_x \) are the precisions required to discretize \( x \) and \( \theta \), respectively. Note that the term \( -n \log \delta_f \) is common across all models, so it can be ignored. However, the term \( -k \log \delta_x \) which involves the dimension of \( \theta \) in the model varies and is influential. According to [6,12], \( \delta_x = 1/\sqrt{n} \) is optimal (see [13] for details), in which case
\[ DL^r(\theta^k) \approx -\log f\left( x^0 | \theta^k \right) - \log \pi(\theta^k) \]
\[ + \frac{k}{2} \log n + \text{constant}. \tag{3} \]

Minimizing this will not lead to MLE even when \( \pi(\theta^k) \) is assumed to be approximately constant. In fact, [12] proceeds further and argues that the correct precision \( \delta_x \) should depend on the Fisher information matrix. This amounts to using a prior which is similar in nature to the Jeffreys’ prior on \( \theta \). Jeffreys prior is an objective choice and thus this approach to MDL can then be considered a default Bayesian approach.

In spite of these desirable properties, however, MDL leads to the HPD estimate of \( \theta \), which is not the usual Bayes estimate. Posterior mean is what is generally preferred, so that the error in estimation has an immediate simple answer in the posterior standard deviation. In summary, therefore, the Bayesian approach doesn’t seem to find attractive solutions in the MDL approach as far as estimation or model fitting is concerned unless the models under consideration are hierarchical having parameters of varying dimension. On the other hand, when such hierarchical models are of interest the inference problem usually involves model selection in addition to model fitting. Thus the possible gains from studying the MDL approach are in the context of model selection as described below.

4. Model Selection Using MDL

Let us recall the Bayesian approach to model selection

Copyright © 2013 SciRes.
and express it in the following form. Let 

$$X = X' = (X_1, \ldots, X_n).$$

Suppose 

$$X' | \theta \sim f(x'| \theta), \theta \in \Theta.$$ Consider testing

$$M_0 : \theta \in \Theta_0 \text{ versus } M_1 : \theta \in \Theta_1,$$ (4)

where \( \Theta_i \subset \Theta \subseteq \mathcal{X}^d \), for some \( d, i = 0, 1, \Theta_0 \cup \Theta_1 = \Theta \) and \( \Theta_0 \cap \Theta_1 = \emptyset \). Let \( \pi \) be a prior on \( \Theta \). Then \( \pi \) can be expressed as

$$\pi(\theta) = \pi_0 g_0(\theta) I(\theta \in \Theta_0) + (1 - \pi_0) g_1(\theta) I(\theta \in \Theta_1)$$ (5)

where \( \pi_0 = P(\Theta_0 | \pi) \) and \( g_0 \) and \( g_1 \) are the conditional densities (with respect to some dominating \( \sigma \)-finite measure) of \( \theta \) under \( M_0 \) and \( M_1 \) respectively. Then

$$P(M_0 | x^n) = \frac{\pi_0 m_0(x^n)}{\pi_0 m_0(x^n) + (1 - \pi_0) m_1(x^n)}$$ (6)

$$= \frac{\pi_0 m_0(x^n)}{m(x^n)},$$ (7)

where

$$m_i(x^n) = \int_{\Theta_i} f(x^n | \theta) g_i(\theta) d\theta, i = 1, 2,$$ (8)

and

$$m(x^n) = \pi_0 m_0(x^n) + (1 - \pi_0) m_1(x^n).$$ (9)

Note that \( m_i \) is simply the marginal or predictive density of \( X^n \) under \( M_i \) and \( m \) is the unconditional predictive density obtained upon averaging \( m_0 \) and \( m_1 \). Consequently, the posterior odds ratio of \( M_0 \) relative to \( M_1 \) is,

$$\frac{P(M_0 | x^n)}{P(M_1 | x^n)} = \frac{\pi_0 m_0(x^n)}{1 - \pi_0} \frac{m_0(x^n)}{m_1(x^n)} = -\frac{\pi_0}{1 - \pi_0} \log B_{01}(x^n),$$ (10)

with \( B_{01} \) denoting the Bayes factor of \( M_0 \) relative to \( M_1 \). When we compare two competing models \( M_0 \) and \( M_1 \), we usually take \( \pi_0 = 1/2 \), and hence settle upon the Bayes factor \( B_{01} \) as the model selection tool. This agrees well with the intuitive notion that the model yielding a better predictive ability must be a better model for the given data.

### 4.1. Mixture MDL and Stochastic Complexity

Let us consider the MDL principle now for model selection between \( M_0 \) and \( M_1 \). Once the conditional prior densities \( g_0 \) and \( g_1 \) are agreed upon, MDL will select that model \( M_i \) which obtains a smaller value for the code length, \(-\log m_i(x^n)\), between the two. This is clearly equivalent to using Bayes factor as the model selection tool, and hence this version of MDL is equivalent to the Bayes factor criterion. In the MDL literature, this version of MDL is known as “mixture MDL”, and is distinguished from the “two-stage MDL” which separately codes the model and the prior. The two-stage MDL can be derived as an approximation to the mixture MDL as discussed later. See [13] for further details and other interesting comparisons and discussion. Let us consider a few examples before examining the need for other versions of MDL.

#### Example 1.

Suppose \( X^n \) is a random sample from \( N(\mu, \sigma^2) \) with known \( \sigma^2 \). We want to test

$$M_0 : \mu = 0 \text{ versus } M_1 : \mu \neq 0.$$ 

Consider the \( N(0, \tau^2) \) prior on \( \mu \) with known \( \tau^2 \) under \( M_1 \). Then the marginal distribution of \( X^n \) is \( N_n(0, \sigma^2 \iota_n) \) under \( M_0 \) and under \( M_1 \) is \( N_n(0, \sigma^2 I_n + \tau^2 11') \). A continuous model and a continuous prior is considered here. Since the precision of the prior parameter is the same across all models upon discretization we will ignore the distinction and proceed with densities. Then, both the Bayes factor criterion and the MDL principle will select \( M_1 \) over \( M_0 \) if and only if

$$-\log m_1(x^n) < -\log m_0(x^n),$$

where \( m_i \) and \( m_0 \) are the corresponding densities. Since we are comparing two logarithms, let us switch to natural logarithms. Then

$$-\log m_0(x^n) = \frac{n}{2} \log (2\pi) + \frac{n}{2} \log \sigma^2 + \frac{1}{2 \sigma^2} \sum x_i^2,$$

and

$$-\log m_1(x^n) = \frac{n}{2} \log (2\pi) + \frac{1}{2} \log [\sigma^2 I_n + \tau^2 11']$$

$$+ \left( x^n' \right) (\sigma^2 I_n + \tau^2 11')^{-1} x^n.$$ 

Noting that

$$\sigma^2 I_n + \tau^2 11' = \sigma^2 I_n + (\tau^2 / \sigma^2) 11' = \sigma^2 \left( I_n + n \tau^2 / \sigma^2 \right)$$

and

$$\left( \sigma^2 I_n + \tau^2 11' \right)^{-1} = \sigma^{-2} \left( I_n + (\tau^2 / \sigma^2) 11' \right)^{-1}$$

we obtain

$$-\log m_i(x^n) = \frac{n}{2} \log (2\pi) + \frac{1}{2} \log \left[ \sigma^2 \left( 1 + n \tau^2 / \sigma^2 \right) \right]$$

$$+ \frac{1}{2 \sigma^2} \sum x_i^2 - \frac{n \tau^2}{2 \sigma^2 \left( n \tau^2 + \sigma^2 \right)} n \bar{x}^2.$$
Therefore $M_1$ is preferred over $M_0$ either by the Bayes factor or by the mixture MDL if and only if
\[
\frac{n}{\sigma^2} \lambda > \frac{n \pi^2 + \sigma^2}{n \pi^2} \log \left( 1 + \frac{n \pi^2}{\sigma^2} \right).
\]

**Example 2.** Let us consider the previous example with unknown $\sigma^2$ now. Suppose the prior on $\sigma^2$ is the default $1/\sigma^2$ under both models. The prior on $\mu$ under $M_1$ is now assumed to depend on $\sigma^2$, i.e., $\mu | \sigma^2 \sim N(0, c \sigma^2)$, where $c > 0$ is assumed to be a known constant for now. Then, provided $x^* \neq 0$, \[m_0(x^*) = \int_0^{\infty} (2\pi \sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \right) \frac{d\sigma^2}{\sigma^2} = \Gamma \left( \frac{n}{2} \right) \pi^{-n/2} \left( \sum_{i=1}^n x_i^2 \right)^{-n/2},\]
and letting $S^2 = \sum_{i=1}^n (x_i - \bar{x})^2$, and $m_i(x|c, \sigma^2)$ denote the marginal density of $\bar{x}$ given $c$ and $\sigma^2$, \[m_i(x^*) = \int_0^{\infty} (2\pi \sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \right) \frac{d\sigma^2}{\sigma^2} = \Gamma \left( \frac{n}{2} \right) \pi^{-n/2} \left( 1 + nc \right)^{-n/2} \left( S^2 + \frac{\bar{x}^2}{c + 1/n} \right)^{-n/2}.
\]
Thus
\[-\log \frac{m_0(x^*)}{m_i(x^*)} = -\log \left( \sum_{i=1}^n x_i^2 \right)^{-n/2} - \log \left( S^2 + \bar{x}^2/(c + 1/n) \right)^{-n/2} (1 + nc)^{-n/2} = -\frac{1}{2} \log (1 + nc) + \frac{n}{2} \log \frac{S^2 + n\bar{x}^2}{S^2 + \bar{x}^2/(c + 1/n)} = -\frac{1}{2} \log (1 + nc) + \frac{n}{2} \log \frac{1 + n\bar{x}^2 / S^2}{1 + n\bar{x}^2 / ((1 + nc) S^2)}.
\]

Therefore, the Bayes factor criterion or the mixture MDL reduces to a criterion which is very similar to that given in the previous example, except that $\sigma^2$ is now replaced by an estimator.

**Example 3. (Jeffreys’ Test)** This is similar to the problem discussed above, except that $\mu | \sigma \sim N(0, \sigma)$, with density \[g_1(\mu|\sigma) = \frac{1}{\sigma \pi} \frac{1}{\left( 1 + \mu^2 / \sigma^2 \right)^{3/2}},\]
the Cauchy prior. The prior on $\sigma$ is the same as before under both models: $\pi_0(\sigma) = 1/\sigma$. This approach suggested by Jeffreys ([14]) is important. It explains how one should proceed when the hypotheses which describe the model selection problem involve only some of the parameters and the remaining parameters are considered to be nuisance parameters. Then Jeffreys suggestion is to employ the same noninformative prior on the nuisance parameters under both models, and a proper prior with low level of information on the parameters of interest. Details on this problem along with this choice of prior can be found in Section 2.7 of [15].

Note that $m_0(x^*)$ is the same as in the previous example, namely \[m_0(x^*) = \int_0^{\infty} (2\pi \sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \right) \frac{d\sigma^2}{\sigma^2} = \Gamma \left( \frac{n}{2} \right) \pi^{-n/2} \left( \sum_{i=1}^n x_i^2 \right)^{-n/2},\]
whereas \[m_i(x^*) = \int_0^{\infty} (2\pi \sigma^2)^{-n/2} \exp \left( -\frac{S^2}{2\sigma^2} \right) m_i(x|c, \sigma^2) \frac{d\sigma^2}{\sigma^2} = \Gamma \left( \frac{n}{2} \right) \pi^{-n/2} \left( 1 + nc \right)^{-n/2} \left( S^2 + \frac{\bar{x}^2}{c + 1/n} \right)^{-n/2}.
\]

No closed form is available for $m_i(x^*)$ in this case. To calculate this one can proceed as follows as indicated in Section 2.7 of [15]. The Cauchy density $g_i(\mu|\sigma)$ can be expressed as a Gamma scale mixture of normals, \[g_i(\mu|\sigma) = \int_0^{\infty} \frac{1}{\sqrt{2\pi} \sigma} \frac{1}{\gamma^{3/2}} e^{-\gamma^2 / 2\sigma^2} \frac{d\gamma}{\sigma \left( 1 + \mu^2 / \gamma^2 \right)^{3/2}} = \frac{1}{\sqrt{2\pi} \gamma^{3/2}} e^{-\gamma^2 / 2\sigma^2} \frac{d\gamma}{\sigma \left( 1 + \mu^2 / \gamma^2 \right)^{3/2}} \]
where $\gamma$ is the mixing Gamma variable. Now one can integrate over $\mu$ and $\sigma$ in closed form to simplify $(m_i(x^*))$. Finally, one has a one-dimensional integral over $\gamma$ left, which can be numerically computed whenever needed.

Now, let us note from the examples discussed above that an efficient computation of $m_i(x^*)$ relies on having an explicit functional form for it. This is generally possible only when a conjugate prior is used as in Examples 1 and 2. For other priors, such as in Example 3, some numerical approximation will have to be employed. Thus we are lead to considering possible approximations to the mixture MDL technique, or equivalently to the Bayes factor, $BF_{01}$. 

Copyright © 2013 SciRes.
From Sections 4.3.1 and 7.1 of [15], assuming that \( f \) and \( g_i \) are smooth functions, we obtain, for large \( n \), the following asymptotic approximation for \( m_i (x^*) \) of equation (8). Let \( k_i \) be the dimension of \( \Theta_i \),
\[
    nh(\theta^k) = -\log f(x^*|\theta^k) \quad \text{and} \quad \Delta_k(\theta^k) \quad \text{denote the Hessian of} \ h, \ i.e.,
\]
\[
    \Delta_k(\theta^k) = \left(-\frac{\partial^2}{\partial \theta_i \partial \theta_j} h(\theta^k)\right)_{i,j=k_i}.
\]

Also, let \( \hat{\theta}^k \) denote either the MLE or the posterior mode. Then
\[
    m_i (x^*) = \int f(x^*|\theta^k) g_i(\theta^k) d\theta^k
\]
\[
= e^{-n \tilde{h}(\theta^k)} \left(2\pi\right)^{k_i/2} n^{-k_i/2} \det(\Delta_k(\theta^k))^{-1/2} \times g_i(\theta^k) \left(1 + O(n^{-1})\right),
\]
so that
\[
    -\log m_i (x^*) = nh(\hat{\theta}^k) - \frac{k_i}{2} \log(2\pi) + \frac{k_i}{2} \log n
\]
\[
+ \frac{1}{2} \log \det(\Delta_k(\hat{\theta}^k)) - \log g_i(\hat{\theta}^k) + O(n^{-1}) \quad (11)
\]
\[
\equiv -\log f(x^*|\hat{\theta}^k) - \frac{k_i}{2} \log(2\pi) + \frac{k_i}{2} \log n
\]
\[
+ \frac{1}{2} \log \det(\Delta_k(\hat{\theta}^k)) - \log g_i(\hat{\theta}^k) + O(n^{-1}).
\]

Ignoring terms that stay bounded as \( n \to \infty \), we obtain the Schwarz criterion ([19]), BIC, for \( M_i \) given by
\[
    \text{BIC}(x^*) = \log f(x^*|\hat{\theta}^k) - \frac{k_i}{2} \log n, \quad (16)
\]
which can be seen to be asymptotically equivalent to SIC.

**Example 4.** [20] discusses a model selection problem for failure time data. The two models considered are exponential and Weibull:
\[
    M_0 : f(x|\lambda, \alpha) = \lambda \exp(-\lambda x), x > 0
\]
versus
\[
    M_i : f(x|\lambda, \alpha) = \lambda x^{\alpha-1} \exp(-\lambda x^\alpha), x > 0,
\]
where \( \lambda > 0 \) and \( \alpha > 0 \). Model selection criterion of Rissanen is SIC as described in Equations (13) and (14). However, in this problem a better approximation is employed for the mixture MDL, the mixture being Jeffreys mixture, i.e., the conditional prior densities under \( M_i \) for \( i = 1, 2 \) are given by
\[
    g_i(\theta^k) = B_i \sqrt{\det(I(\theta^k))}, \theta^k \in \Gamma_i,
\]
where \( \Gamma_i \) is a compact subset of the relevant parameter space \( \Theta_i \) and
\[
    B_i = \int_{\Gamma_i} \sqrt{\det(I(\theta^k))} d\theta^k.
\]
Consequently, it follows that,
\[
    m_i (x^*) = B_i \int_{\Gamma_i} f(x^*|\hat{\theta}^k) \sqrt{\det(I(\theta^k))} d\theta^k
\]
\[
= B_i f(x^*|\hat{\theta}^k) \left(\frac{n}{2\pi}\right)^{-k_i/2} + O(n^{-1}),
\]
where \( \hat{\theta}^k \) is the MLE of \( \theta^k \) under \( M_i \). This yields,
\[
    -\log m_i (x^*) = -\log f(x^*|\hat{\theta}^k) + \frac{k_i}{2} \log \frac{n}{2\pi}
\]
\[
+ \log \int_{\Gamma_i} \left(\sqrt{\det(I(\theta^k))}\right)^{1/2} d\theta^k + O\left(\frac{1}{n}\right).
\]

Compare this with (15) and note that the term involving \( \det(I(\hat{\theta}^k)) \) vanishes.

We would like to note here that many authors [21,22] define the MDL estimate to be the same as the HPD estimate with respect to the Jeffreys’ prior restricted to some compact set \( K \) where its integral is finite:
\[
    \sqrt{\frac{\det(I(\theta))}{\int_{K} \det(I(u)) du}}, \text{for} \ \theta \in K,
\]
which is the stochastic complexity approach advocated.
above.

It must be emphasized that proper priors are being employed to derive the SIC criterion, and hence indeterminacy and inconsistency problems faced by techniques employing improper priors are not a difficulty in this approach. Moreover, this approach can be viewed as an implementable approximation to an objective Bayesian solution.

### 4.2. Two-Stage MDL

Now consider the two-stage MDL which codes the prior and the likelihood separately and adds the two description lengths. This approach is therefore similar to estimating the parameter $\theta$ with the HPD estimate when there is an informative prior, or with the MLE, but the resulting minimum description length does have interesting features. To see when and how this approach approximates the above mentioned model selection criterion, let us look at some of the specific details in the two stages of coding. See [12,13] for further details. Again, recall the setup in (4) and (5).

**Stage 1.** Let $\hat{\theta}_j^k$ be an estimate of $\theta_j^k$ such as the posterior mean, HPD or MLE under $M_j$. This needs to be coded. Consider the prior density $g_j(\theta_j^k)$ conditional on $M_j$ being true. Usually MDL would choose a uniform density. Restrict $\theta$ to a large compact subset of the parameter space and discretize it as discussed in Section 3 with a precision of $\delta_j = 1/\sqrt{n}$. Then the code-length required for coding $\hat{\theta}_j^k$ is

$$L(\hat{\theta}_j^k) = -\log g_j(\hat{\theta}_j^k) + \frac{k_j}{2} \log n. \quad (17)$$

**Stage 2.** Now the data $x^n$ is coded using the model density $f(x^n|\hat{\theta}_j^k)$. Discretization may again be needed, say with precision $\delta_j$. Thus the description length for coding $x^n$ will be

$$-\log f(x^n|\hat{\theta}_j^k) - n \log \delta_j. \quad (18)$$

Summing these two codelengths, therefore, we obtain a total description length of

$$-\log f(x^n|\hat{\theta}_j^k) - n \log \delta_j - \log g_j(\hat{\theta}_j^k) + \frac{k_j}{2} \log n. \quad (19)$$

Since the second term above, $-n \log \delta_j$, is constant over both $M_0$ and $M_1$, and the third term stays bounded as $n$ increases, these two terms are dropped from the MDL two-stage coding criterion. Thus, for regular parametric models, the two-stage MDL simplifies to the same criterion (for $M_j$) as BIC, namely,

$$-\log f(x^n|\hat{\theta}_j^k) + \frac{k_j}{2} \log n. \quad (20)$$

In more complicated model selection problems, the two-stage MDL will involve further steps and may differ from BIC.

It may also be seen upon comparing (19) with (15) that the performance of SIC based MDL should be superior to the simplified two-stage MDL for moderate $n$ since SIC uses a better precision for coding the parameter, namely, one based on the Fisher information.

### 5. Regression and Function Estimation

Model selection is an important part of parametric and nonparametric regression and smoothing. Variable selection problems in multiple linear regression, order of the spline to fit and wavelet thresholding are some such areas. We will briefly consider these problems to see how MDL methods can provide computationally attractive approximations to the respective Bayesian solutions.

#### 5.1. Variable Selection in Linear Regression

Variable selection is an important and well studied problem in the context of normal linear models. Literature includes [23-32]. We will only touch upon this area with the specific intention of examining useful and computationally attractive approximations to some of the Bayesian methods.

Suppose we have an observation vector $y^n$ on a response variable $Y$ and also measurements $x^n_1, x^n_2, \ldots, x^n_M$ on a set of potential explanatory variables (or regressors). Following [13], we associate with each regressor $x_j$, a binary variable $\gamma_j$. Then the set of available linear models is

$$y^n = \sum_{j:j\gamma_j=1} \beta_j x^n_j + \epsilon^n, \quad (21)$$

where $\epsilon^n \sim N_n(0, \sigma^2 I_n)$. Note that $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_M)$ is, then, a Bernoulli sequence associated with the set of regression coefficients, $\beta^M = (\beta_1, \beta_2, \ldots, \beta_M)$ also. Let $\beta_j$ denote the vector of non-zero regression coefficients corresponding to $\gamma_j$ and $X_j$ the corresponding design matrix, which results in the model

$$y^n = X_j \beta_j + \epsilon^n.$$

Selecting the best model, then, is actually an estimation problem, i.e., find the HPD estimate of $\gamma$ starting with a prior $\pi_1$ on $\gamma$ and a prior $\pi_2$ on $\left(\beta_j, \sigma^2\right)$ given $\gamma$. The two-stage MDL, which is the simplest, uses the criterion of minimizing

$$DL(y^n|\gamma) = -\log f(y^n|\hat{\beta}_j, \hat{\sigma}_j^2, X_j, \gamma) - \log \pi_1(\gamma). \quad (22)$$

MLE for $\beta_j$ and $\sigma^2$ given $\gamma$ are easily available:
\[ \hat{\beta}_f = (X_f'X_f)^{-1}X_f'y, \]
\[ \hat{\sigma}^2 = \text{RSS}_f/n = \|y - X_f\hat{\beta}_f\|^2/n. \] (23)

Consider the uniform prior on \( \gamma \), all \( 2^M \) values receiving the same weight \( 1/2^M \). Using these, we can re-write the MDL criterion as the one which minimizes (as in Example 2)
\[ \frac{n}{2} \log \text{RSS} + \frac{k}{2} \log n, \] (24)
where \( k \) is the number of \( \{\gamma_j = 1\} \).

We can also derive the mixture MDL or stochastic complexity of a given model \( \gamma \). If \( g(\beta, \sigma^2|\gamma) \) is the prior density under \( \pi_\gamma \),
\[ m(y|\gamma) = \int f(y|\beta, \sigma^2, X, \gamma) g(\beta, \sigma^2|\gamma) \, d\beta, d\sigma^2. \] (25)

Applying (13), (14) after evaluating the information matrix of the parameters \( \{\beta, \sigma^2\} \) and ignoring terms that are irrelevant for model selection, one obtains (see [13]),
\[ \text{SIC}(\gamma) = \frac{n-k-2}{2} \log \text{RSS}_\gamma \]
\[ + \frac{k}{2} \log n + \frac{1}{2} \log \det(X'_\gamma X_\gamma). \] (26)

If \( g \) is chosen to be the conjugate prior density, then the marginal density \( m(y|\gamma) \) can be explicitly derived. Details on this and further simplifications obtained upon using Zellner’s g-prior can be found in [13]. (See also [33,34].)

This method is only useful if one is interested in comparing a few of these models, arising out of some pre-specified subsets. Comparing all \( 2^M \) models is not a computationally viable option for even moderate values of \( M \), since for each model, \( \gamma \), one has to compute the corresponding \( \hat{\beta}_\gamma \) and \( \text{RSS}_\gamma \).

We are more interested in a different problem, namely, whether an extra regressor should be added to an already determined model. This is the idea behind the step-wise regression, forward selection method. In this set-up, the model comparison problem can be stated as comparing
\[ M_0 : y^n = \sum_{j=1}^{k} \beta_j x^n_j + e^n \]
versus
\[ M_1 : y^n = \sum_{j=1}^{k+1} \beta_j x^n_j + e^n. \]

This is actually a model building method, so we assume that \( x^n_0 = 1^n \), and hence \( \hat{\beta}_0 \) is the intercept which gives the starting model. Then we decide whether this model needs to be expanded by adding additional regressors. Thus, at step \( k \), we have an existing model with regressors \( x_1, \ldots, x_k \) and we fix \( x_{k+1} \) to be one of the remaining \( M-k \) regressors as the candidate for possible selection. Now the two-stage MDL approach is straight forward. From (22) and (24), we note that \( x_{k+1} \) is to be selected if and only if
\[ DL(k+1) - DL(k) = \frac{n}{2} \log \left( \frac{\text{RSS}_{k+1}}{\text{RSS}_k} \right) + \frac{1}{2} \log n < 0, \] (27)
where \( DL(j) \) is the description length of the model with regressors \( x_1, \ldots, x_j \) and \( \text{RSS}_j \) is its residual sum of squares as given in (23). A closer look at (27) reveals certain interesting facts. We need the following additional notations involving design matrices and the corresponding projection matrices. We assume that the required matrix inverses exist.
\[ X_{(j)} = (x^n_1 x^n_2 \ldots x^n_j), \quad p(j) = X_{(j)} (X_{(j)}'X_{(j)})^{-1} X_{(j)}'. \]

Then we note the following result which may be found, for example, in [35].
\[ (X_{(j+1)}'X_{(j+1)})^{-1} = \begin{pmatrix} X_{(j)}'X_{(j)} & X_{(j)}'x_{k+1} \\ x_{k+1}'X_{(j)} & x_{k+1}'x_{k+1} \end{pmatrix}^{-1} = \begin{pmatrix} (X_{(j)}'X_{(j)})^{-1} + \nu uu' & -\nu' \\ -\nu' & \nu \end{pmatrix}, \] (28)

where
\[ u = (X_{(j)}'X_{(j)})^{-1}x_{k+1}, \]
and \( \nu = (x_{k+1}'(I_n - p(j))x_{k+1})^{-1} \).

It then follows that
\[ p_{(k+1)} = p(j) + \nu x_{k+1} x_{k+1}' - x_{k+1}' x_{k+1} p(j) \]
and hence
\[ \text{RSS}_{k+1} - \text{RSS}_k = y'(p_{(k+1)} - p(j)) y \]
\[ = \nu [y' (I_n - p(j)) x_{k+1}]^2, \]
and
\[ \frac{\text{RSS}_{k+1} - \text{RSS}_k}{\text{RSS}_k} = \frac{[y'(I_n - p(j)) x_{k+1}]^2}{[x_{k+1}' (I_n - p(j)) x_{k+1}]^2}, \]
\[ = r^2_{y',x_{k+1}}(1,2,\ldots,k), \]
where \( r \) is simply the partial correlation coefficient between \( y \) and \( x_{k+1} \), conditional on \( x_1, \ldots, x_k \). Substituting these in (27), we see that

\[
DL(k+1) - DL(k) = \frac{n}{2} \log\left\{ \frac{RSS_{k+1}}{RSS_k} + \frac{1}{2} \log n \right\}
= \frac{n}{2} \log\left(1 - \frac{RSS_k - RSS_{k+1}}{RSS_k}\right) + \frac{1}{2} \log n
= \frac{n}{2} \log\left(1 - r^2_{y,x_{k+1}(1,2,\ldots,k)}\right) + \frac{1}{2} \log n.
\]

(29)

Therefore,

\[
DL(k+1) < DL(k) \text{ if and only if } \frac{n}{2} \log\left(1 - r^2_{y,x_{k+1}(1,2,\ldots,k)}\right) < -\frac{1}{2} \log n \text{ if and only if } r^2_{y,x_{k+1}(1,2,\ldots,k)} > 1 - n^{-1/n}.
\]

This method does have some appeal, in that at each step, it tries to select that variable which has the largest partial correlation with the response (conditional on the variables which are already in the model), just like the step-wise regression method. However, unlike the step-wise regression method it does not require any stopping rule to decide whether the candidate should be added. It relies on the magnitude of the partial correlation instead.

One can also apply the stochastic complexity criterion given in (26) above. Then we obtain,

\[
\text{SIC}(k+1) - \text{SIC}(k) = \frac{n-k-2}{2} \log\left(1 - r^2_{y,x_{k+1}(1,2,\ldots,k)}\right) + \frac{1}{2} \log n
+ \frac{1}{2} \log \left\{ \frac{\det (X'_{(k+1)}X_{(k+1)})}{\det (X'_{(k)}X_{(k)})} \right\}
= \frac{n-k-2}{2} \log\left(1 - r^2_{y,x_{k+1}(1,2,\ldots,k)}\right) + \frac{1}{2} \log n
+ \frac{1}{2} \log \left\{ \frac{\det (X'_{(k)}X_{(k)})}{\det (X'_{(k+1)}X_{(k+1)})} \right\}
= \frac{n-k-2}{2} \log\left(1 - r^2_{y,x_{k+1}(1,2,\ldots,k)}\right)
+ \frac{1}{2} \log n - \frac{1}{2} \log \left( \frac{y' (I_n - \hat{P}_{(k+1)}) y}{x'_{k+1} (I_n - \hat{P}_{(k)}) x_{k+1}} \right).
\]

(31)

which is related to the step-wise regression approach, but uses more information than just the partial correlation.

A full-fledged Bayesian approach using the g-prior can also be implemented as shown below. Note that

\[
m_0\left(y^*\right) = \int \int f\left(y\left|\beta^*, X_{(k)}, \sigma^2\right\rangle g_0\left(\beta^*, \sigma^2\right) d\beta^* d\sigma^2,
\]

and

\[
m_1\left(y^*\right) = \int \int f\left(y\left|\beta^{k+1}, X_{(k+1)}, \sigma^2\right\rangle g_1\left(\beta^{k+1}, \sigma^2\right) d\beta^{k+1} d\sigma^2
\]

where \( g_0 \) and \( g_1 \), respectively, are the prior densities under \( M_0 \) and \( M_1 \). Taking these priors to be g-priors, namely,

\[
\beta^*|c, \sigma^2 \sim N\left(0, c\sigma^2 \left(X'_{(j)}X_{(j)}\right)^{-1}\right), j = k, k+1,
\]

along with the density \( 1/\sigma^2 \) for \( \sigma^2 \), a (proper prior) density \( \pi(\cdot) \) for the hyperparameter \( c \), we obtain

\[
BF_{01} = \frac{\int m_0\left(y^*|c\right)\pi(c) dc}{\int m_1\left(y^*|c\right)\pi(c) dc},
\]

(33)

where

\[
m_0\left(y^*|c\right) = a_n \left(1+c\right)^{-n/2}\left[ RSS_n + \frac{1}{1+c} \hat{P}_{(k)}X'_{(k)}X_{(k)}\hat{P}_{(k)}\right]^{-n/2},
\]

(34)

The one-dimensional integrals in (33), however, cannot be obtained in closed form. One could also approximate \( BF_{01} \) with \( m_0\left(y^*|\hat{c}_k\right)/m_1\left(y^*|\hat{c}_k\right) \), where \( \hat{c} \) are the ML-II (cf. [36]) estimates of \( c \). See [13] for details.

**Example 5.** We illustrate the MDL approach to step-wise regression by applying it to the Iowa-corn-yield-data (see [37,38]). We have not included “year” as a regressor (which is a proxy for technological advance) and instead have considered only the weather-related regressors.

In this data set the variables are: \( X_1 = \text{Year} \), 1 denoting 1930, \( X_2 = \text{Pre-season precipitation} \), \( X_3 = \text{May temperature} \), \( X_4 = \text{June rain} \), \( X_5 = \text{June temperature} \), \( X_6 = \text{July rain} \), \( X_7 = \text{July temperature} \), \( X_8 = \text{August rain} \), \( X_9 = \text{August temperature} \), and \( Y = X_{10} = \text{Corn Yield} \).

As mentioned earlier, we always keep the intercept and check whether this regression should be enlarged by
adding more regressors. We first apply the Two-stage MDL criterion. From (30), at step \( k \), we consider only those regressors (which are not already in the model and) which satisfy \( r^2_{j,k} > 1 - \frac{1}{n^2} \) (=0.1005 in this example). From this set we pick the one with the largest \( r^2_{j,k} \), The values of \( r^2_{j,k} \) for the relevant steps are listed below.

<table>
<thead>
<tr>
<th>step</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
<th>( X_5 )</th>
<th>( X_6 )</th>
<th>( X_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.037</td>
<td>0.010</td>
<td>0.021</td>
<td>0.022</td>
<td>0.338</td>
<td>0.336</td>
<td>0.044</td>
</tr>
<tr>
<td>2</td>
<td>0.066</td>
<td>0.034</td>
<td>0.001</td>
<td>0.015</td>
<td>0.162</td>
<td>0.060</td>
<td>0.018</td>
</tr>
<tr>
<td>3</td>
<td>0.016</td>
<td>0.011</td>
<td>0.022</td>
<td>0.000</td>
<td>--</td>
<td>0.050</td>
<td>0.004</td>
</tr>
</tbody>
</table>

According to our procedure we select \( X_6 \) first, followed by \( X_7 \) and the selection ends there.

We consider the SIC criterion next. From (31), at step \( k \), we pick the regressor \( X_j \) with the largest value for

\[
v(j,k) = r^2_{j,k} + \left[ \frac{\gamma^2(I_n - P^k)}{n\gamma^2(I_n - P^k)} \right] - 1,
\]

provided it is positive. The values of \( v(j,k) \) for the relevant steps are given below.

<table>
<thead>
<tr>
<th>step</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
<th>( X_5 )</th>
<th>( X_6 )</th>
<th>( X_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.002</td>
<td>-0.020</td>
<td>0.043</td>
<td>-0.001</td>
<td>0.371</td>
<td>0.319</td>
<td>0.071</td>
</tr>
<tr>
<td>2</td>
<td>0.013</td>
<td>-0.011</td>
<td>0.012</td>
<td>-0.022</td>
<td>0.141</td>
<td>0.073</td>
<td>-0.002</td>
</tr>
<tr>
<td>3</td>
<td>-0.040</td>
<td>-0.040</td>
<td>0.032</td>
<td>-0.042</td>
<td>--</td>
<td>0.058</td>
<td>-0.022</td>
</tr>
<tr>
<td>4</td>
<td>-0.050</td>
<td>-0.042</td>
<td>0.027</td>
<td>-0.039</td>
<td>--</td>
<td>--</td>
<td>-0.023</td>
</tr>
<tr>
<td>5</td>
<td>-0.043</td>
<td>-0.027</td>
<td>--</td>
<td>-0.028</td>
<td>--</td>
<td>--</td>
<td>-0.025</td>
</tr>
</tbody>
</table>

According to SIC our order of selection is \( X_6, X_7, X_8, X_9, X_4 \).

### 5.2. Wavelet Thresholding

Consider the nonparametric regression problem where we have the following model for the noisy observations \( \nu = (\nu_1, \nu_2, \ldots, \nu_n) \):

\[
\nu_j = s(x_j) + \epsilon_j, j = 1, \ldots, n, \text{and } x_j \in \mathcal{X},
\]

where \( \epsilon_j \) are i.i.d. \( \mathcal{N}(0, \sigma^2) \) errors with unknown error variance \( \sigma^2 \), and \( s \) is a function (or signal) defined on some interval \( \mathcal{X} \). Assuming \( s \) is a smooth function satisfying certain regularities (see [15,39,40]), we have the wavelet decomposition of \( s \):

\[
s(x) = \sum \beta_i \psi_i(x),
\]

where \( \psi_i \) are the wavelets and \( \beta = (\beta_1, \beta_2, \ldots) \) is the corresponding vector of wavelet coefficients. We assume that the normally infinite sum in (36) can be taken to be a finite sum (or at least a very good approximation) as indicated in [39].

Upon applying the discrete wavelet transform (DWT) to \( \nu \), we get the estimated wavelet coefficients, \( x^* = W\nu \). Consider now the equivalent model:

\[
x^* = \beta + \epsilon,
\]

where \( \epsilon \sim \mathcal{N}_n(0, (\sigma^2/n)I_n) \).

The model selection problem here involves determining the number of non-zero wavelet coefficients:

\[
M_0 : \text{number of non-zero } \beta_j \text{ is } k_0
\]

versus

\[
M_i : \text{number of non-zero } \beta_j \text{ is } k_i
\]

where \( k_i < n, i = 0, 1 \) are the number of wavelet coefficients of interest.

The prior distribution on the non-zero \( \beta_j \) is assumed to be i.i.d. \( \mathcal{N}(0, r^2), 1 \leq j \leq k_i \) under \( M_i, i = 0, 1 \).

Since we have not identified the locations (indices) of the non-zero wavelet coefficients, \( \beta_j \), we proceed as follows to describe the prior structure. With each \( \beta_j \) we associate a binary variable \( \gamma_j \) as in [41] for wavelet regression or as in [13] for variable selection in regression. Then \( \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n) \) is a Bernoulli sequence associated with the set of regression coefficients, \( \beta = (\beta_1, \beta_2, \ldots, \beta_n) \). Let

\[
\Gamma_i = \left\{ \gamma : \sum_{j=1}^n \gamma_j = k_i \right\}.
\]

Finally, we let \( z_j, 1 \leq j \leq n \) be i.i.d. \( \mathcal{N}(0, r^2) \) ( \( g \) be the corresponding joint density), and define the following structure under \( M_i \):

\[
\pi(\gamma) = \begin{cases} \frac{n!}{k_i^{k_i}(n-k_i)^{n-k_i}} & \text{for all } \gamma : \sum_{j=1}^n \gamma_j = k_i; \\ 0 & \text{otherwise}. \end{cases}
\]

(37)

\[
\beta | \gamma, k_i : \beta_j = \gamma_j z_j, 1 \leq j \leq n.
\]

(38)

\[
f(x^* | \beta, \sigma^2) = (2\pi\sigma^2/n)^{-n/2} \exp \left[ -\frac{n}{2\sigma^2} \sum_{j=1}^n (x_j - \gamma_j z_j)^2 \right].
\]

(39)

The nuisance parameter \( \sigma^2 \) which is common under both models is given the prior density \( 1/\sigma^2 \). Then it follows that
\[ m_i(x^o) = m_i(x^o|k_i) \]
\[ = \sum_{\gamma \in \Gamma} \frac{1}{n} \left[ \int_{x \in x^o} f(x^o|z, \gamma, k_i, \sigma^2) g(z) dz \right] d\sigma^2 \]
\[ = \sum_{\gamma \in \Gamma} \frac{1}{n} \int_{x \in x^o} (2\pi\sigma^2/n)^{-\sigma^2} \exp\left( -\frac{n}{2\sigma^2} \sum_{j=1}^{n}(x_j - \gamma_j z_j)^2 \right) \times (2\pi\sigma^2)^{-\sigma^2} \exp\left( -\frac{1}{2\sigma^2} \sum_{j=1}^{n} z_j^2 \right) dz \]  

(40)

Note that only those \( z_j \) for which \( \gamma_j = 1 \) appear in the integral above.

The Two-stage MDL approach is clearly the easiest to take in this problem. As described earlier, it approximates \( m_i(x^o) \) by coding the prior and the likelihood (both evaluated at an estimate) separately and sums the codelengths to obtain the description length. In this case, discretizing \( \beta_j \) and \( \sigma^2 \) to a precision of \( 1/\sqrt{n} \) and ignoring terms that stay bounded as \( n \) increases, this amounts to

\[ DL(x^o|k_i) \]
\[ = -\log \pi(\gamma|k_i) - \log \pi(\hat{\beta}|\gamma, k_i) \]
\[ -\log \pi(\sigma^2) - \log f(x^o|\hat{\beta}, \hat{\sigma}^2) \]
\[ \approx \log \left( \frac{n}{k_i} \right) + \frac{1}{2\tau^2} \sum_{j=1}^{n} \hat{\beta}_j^2 \]
\[ + \frac{n}{2} \log \left( \frac{2\pi\hat{\sigma}^2}{n} \right) + \frac{n}{2\sigma^2} \sum_{j=1}^{n} (x_j - \hat{\beta}_j)^2 \]
\[ \approx k_i \log n + \frac{k_i}{2} \log n + \frac{n}{2} \log \left( \frac{2\pi\hat{\sigma}^2}{n} \right) \]  

(41)

where the first term is obtained from Stirling’s approximation, the second term \( k_i/2 \log n \) is for coding the \( k_i \) non-zero \( \beta_j \)’s and \( \hat{\sigma}^2 \) is an estimate such as the MLE. On the other hand, computing SIC or \( m_i(x^o) \) is not an impossible task either. In fact, to integrate out the \( z_j \) in \( m_i(x^o) \) of Equation (40) we argue as follows.

\[ X_j \mid (\gamma_j = 0, \sigma^2) \sim N(0, \sigma^2/n), \]  

and \( \tau^2 = c \sigma^2 \), and integrate out \( \sigma^2 \) also. This leaves us with the following expression where we have a sum over \( \gamma \).

\[ m_i(x^o) \]
\[ = \sum_{\gamma \in \Gamma} \frac{1}{n} \left[ \int_{x \in x^o} (2\pi\sigma^2/n)^{-\sigma^2} \exp\left( -\frac{n}{2\sigma^2} \sum_{j=1}^{n}(x_j - \gamma_j z_j)^2 \right) \times (2\pi\sigma^2)^{-\sigma^2} \exp\left( -\frac{1}{2\sigma^2} \sum_{j=1}^{n} z_j^2 \right) dz \right] d\sigma^2 \]  

(42)

The term \( \left( \sum_{j=1}^{n}(1-\gamma_j) x_j^2 + \frac{1}{1+nc} \sum_{j=1}^{n} y_j x_j^2 \right)^{-n/2} \) is interesting. Most of the contribution to this sum is expected from \( \sum_{j=1}^{n}(1-\gamma_j) x_j^2 \) with \( \gamma_j = 1 \) corresponding to the largest \( k_i \) of the \( x_j \), which yields the MLE of \( \sigma^2 \) upon normalization. The Bayes estimate, on the other hand, will arise from a weighted average of all the sums, with weights depending on the posterior probabilities of the corresponding \( \gamma \). As is clear, weighted average over the space \( \Gamma_j \) is computationally very intensive when \( n \) and \( k_i \) are large. An appropriate approximation is indeed necessary, and MDL is important in that sense.

Even though we have justified the two-stage MDL for wavelet thresholding by showing that it is an appro-
ximation to a mixture MDL corresponding to a certain prior, a few questions related to this prior remain. First of all, the prior assumption that $x_i$ are i.i.d. $N(0, \tau^2)$ is unreasonable; wavelet coefficients corresponding to wavelets at different levels of resolution must be modeled with different variances. Specifically they should decrease as the resolution level increases to indicate their decreasing importance (see [40,42,43]). Secondly, wavelet coefficients tend to cluster according to resolution levels (see [44]), so instead of independent normal priors, a multivariate normal prior with an appropriate dependence structure must be employed. These modifications can be easily implemented in the Bayesian approach, except that the resulting computational requirements may be substantial.

5.3. Change Point Problem

We shall now consider MDL methods for a problem which attempts to decide whether there is a change-point in a given time series data. We use the data on British road casualties available in [45], which examines the effects on casualty rates of the seat belt law introduced on 31 January 1983 in Great Britain.

We follow the approach of [46]. Let $Y_1, Y_2, \ldots, Y_n$ be independent Poisson counts with $Y_i \sim \text{Poisson}(\lambda_i)$. $\lambda_i$ are a priori considered related, and a joint multivariate normal prior distribution on their logarithm is assumed. Specifically, let $v_i = \log(\lambda_i)$ be the $i$th element of $v$ and suppose

$$v | \mu \sim N_n(\mu, \delta^2 \Gamma).$$

We model the change-point as the model selection problem:

$$M_0 : \mu_1 = \mu_2 = \cdots = \mu_n = \xi_0,$$

$$\Gamma = (1 - \rho)I_n + \rho \mathbf{1}_n \mathbf{1}_n' = \Gamma_0,$$

versus

$$M_1 : \mu_1 = \mu_2 = \cdots = \mu_n = \xi_0,$$

$$\mu_{n+1} = \mu_{n+2} = \cdots = \mu_\infty = \xi_1,$$

$$\Gamma = (1 - \rho)I_n + \rho \mathbf{1}_n \mathbf{1}_n' = \Gamma_1,$$

where $n_1$ is the possible change-point. We further let $\xi_0, \xi_1$ be i.i.d. $N(0, \tau^2)$. Note that $\delta^2, \rho$ and $\tau^2$ are hyperparameters.

First, we approximate the likelihood function assuming $y_i > 0$ as follows.

$$f(y | v) = \exp \left( -\sum_{i=1}^n \{e^{y_i} - v_i y_i\} \right) / \prod_{i=1}^n y_i!$$

$$= \exp \left( -\sum_{i=1}^n e^{y_i} \log y_i + \sum_{i=1}^n y_i \{v_i - \log y_i\} + \sum_{i=1}^n u_i \log y_i \right)$$

$$= \exp \left( \sum_{i=1}^n v_i \{v_i - e^{y_i}\} \right) \exp \left( \sum_{i=1}^n y_i \log y_i \right),$$

where $v_i = v_i - \log y_i$. Expanding (43) about $0$, its maximum, in Taylor series and ignoring higher order terms in $v_i$, we obtain

$$f(y | v) \approx \exp \left( \sum_{i=1}^n v_i \{\log y_i - 1\} \right)$$

$$= \exp \left( -\frac{1}{2} \sum_{i=1}^n y_i (\log y_i - v_i^2) \right).$$

What is appealing and useful about (44) is that it is proportional to the multivariate normal likelihood function (for $v$) with mean vector $x$ and covariance matrix $W$ where

$$x = (x_1, \ldots, x_n) = (\log y_1, \ldots, \log y_n),$$

and

$$W = \text{diag}(w_1, \ldots, w_\infty) = \text{diag}(1/y_1, \ldots, 1/y_n).$$

Thus hierarchical Bayesian analysis of multivariate normal models is applicable (see [15,36,47]). We note that the hyper-parameters $\rho$ and $\tau^2$ do not have substantial influence and hence treat them as fixed constants (to be chosen based on some sensitivity analysis) in the following discussion. Consequently, denoting by $g_0$ and $g_1$ the respective prior densities under $M_0$ and $M_1$,

$$m_0(y^* | \rho, \tau^2) = \int f(y^* | v) g_0(v, \xi_0, \delta^2) dv d\xi_0 d\delta^2$$

$$= \exp \left( \sum_{i=1}^n v_i \{\log y_i - 1\} \right) \{2\pi\}^{n/2} \text{det}(W)^{1/2} \times \left( (2\pi) \frac{1}{2} \int \frac{1}{2} \left( x + y \right)^{n-1} (x+y) g_0(v, \xi_0, \delta^2) dv d\xi_0 d\delta^2 \right)$$

$$= \frac{\exp \left( \sum_{i=1}^n v_i \{\log y_i - 1\} \right) \{2\pi\}^{n/2} \text{det}(W)^{1/2}}{\left( (2\pi) \frac{1}{2} \int \frac{1}{2} \left( x + y \right)^{-1} (x+y) g_0(v, \xi_0, \delta^2) dv d\xi_0 d\delta^2 \right)^{1/2}}$$

$$\times \left( (2\pi) \frac{1}{2} \int \frac{1}{2} \left( x + y \right)^{-1} (x+y) g_0(v, \xi_0, \delta^2) dv d\xi_0 d\delta^2 \right)^{1/2}.$$
Now, from multivariate normal theory, observe that
\[
\int (2\pi)^{-n/2} |W|^{-1/2} \exp \left\{ -(x-y)^T W^{-1} (x-y) \right\} \delta^2 \left\{ (1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n^T \right\}^{-1/2} \delta^2 \delta^2 dx \delta^2 dy \delta^2 dz
\]
\[
\times \exp \left\{ -\frac{1}{2} (\nu - \xi_0)^T (\delta^2 \left\{ (1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n^T \right\}^{-1} (\nu - \xi_0) \right\} dv
\]
\[
= (2\pi)^{-n/2} |W + \delta^2 \left\{ (1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n^T \right\} + \tau^2 \mathbf{1}_n \mathbf{1}_n^T|^{-1/2} \times \exp \left\{ -\frac{1}{2} (x - \xi_0)^T (W + \delta^2 \left\{ (1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n^T \right\}^{-1} (x - \xi_0) \right\},
\]
and subsequently that, \( \xi_0 \) can be integrated out as in Example 2. Thus,
\[
m_0 \left( y^0 \mid \rho, \tau^2 \right) \approx \exp \left( \sum_{i=1}^n y_i \left\{ \log y_i - 1 \right\} \right) \times (2\pi)^{-n/2} |W + \delta^2 \left\{ (1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n^T \right\} + \tau^2 \mathbf{1}_n \mathbf{1}_n^T|^{-1/2} \times \exp \left\{ -\frac{1}{2} \left[ \nu - \xi_0 \right]^T \left( W + \delta^2 \left\{ (1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n^T \right\} + \tau^2 \mathbf{1}_n \mathbf{1}_n^T \right]^{-1} \left[ \nu - \xi_0 \right] \right\} g_0 \left( \delta^2 \right) d\delta^2
\]
Expression (47) is not available, in general, in closed form. Approaching it from the MDL technique, we look for a subsequent approximation employing an ML-II type estimator (cf. [36]) for \( \delta^2 \). Again, from (47), an ML-II likelihood for \( \delta^2 \) is given by \( \delta^2 \) which maximizes
\[
\frac{1}{2} \left[ \nu - \xi_0 \right]^T \left( W \right) \left[ \nu - \xi_0 \right] + \frac{1}{2} \left[ \nu - \xi_0 \right]^T \left( \frac{\nu - \xi_0}{\delta^2} \right) \left[ \nu - \xi_0 \right] \delta^2 x \delta^2 y \delta^2 z
\]
For fixed \( \rho \) and \( \tau^2 \) this involves only examining a smooth function of a single variable, \( \delta^2 \), which is a simple computational task.

We proceed exactly as above to derive \( m_1 \left( y^0 \mid \rho, \tau^2 \right) \) also. Partition the required vectors and matrices as follows.
\[
\nu^0 = \left( \nu^0, \nu^1 \right), \nu^0 = \left( y^0, y^1 \right), x^0 = \left( x^0, x^1 \right),
\]
\[
w^0 = \left( w^0, w^1 \right), W = \begin{pmatrix} W_{11} & 0 \\ 0 & W_{22} \end{pmatrix},
\]
where \( W_{11} = \text{diag} \left( w^0 \right) \) and \( W_{22} = \text{diag} \left( w^1 \right) \).

Then we have
\[
m_1 \left( y^0 \mid \rho, \tau^2 \right) = \int \left( y^0 \right) g_1 \left( \nu^0, \xi_0, \xi_1, \delta^2 \right) d\nu^0 d\xi_0 d\xi_1 d\delta^2
\]
\[
\approx \exp \left( \sum_{i=1}^n y_i \left\{ \log y_i - 1 \right\} \right) \left( 2\pi \right)^{-n/2} \det(W_1) \left( 2\pi \right)^{-n/2} |W_1 + \delta^2 \left\{ (1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n^T \right\} + \tau^2 \mathbf{1}_n \mathbf{1}_n^T|^{-1/2} \times \exp \left\{ -\frac{1}{2} \left[ \nu^0 - \xi_0 \right]^T \left( W_{11} + \delta^2 \left\{ (1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n^T \right\} + \tau^2 \mathbf{1}_n \mathbf{1}_n^T \right]^{-1} \left[ \nu^0 - \xi_0 \right] \right\} g_1 \left( \nu^0, \xi_0, \xi_1, \delta^2 \right) d\nu^0 d\xi_0 d\xi_1 d\delta^2
\]

(48)
As before, the MDL technique involves deriving the ML-II estimator of $\delta^2$ from $m_i(\nu; \delta^2, \rho, \tau^2)$, for fixed $\rho$ and $\tau^2$. Obtaining $\hat{\delta}^2_i$ (for fixed $\rho$ and $\tau^2$) which maximizes $m_i(\nu; \delta^2, \rho, \tau^2)$ is very similar to that for $\hat{\delta}^0_i$.

We have applied this technique to analyze the British Road Casualties data. Figures 1(a) and (b) show $\log(m_0(\delta^2, \rho, \tau^2))$ and $\log(m_1(\delta^2, \rho, \tau^2))$ as a function of $\delta^2$ for $\rho = 0.1$ and $\tau^2 = 5$, for the LGV and HGV data, respectively. As mentioned previously, $\rho$ and $\tau^2$ do not seem to play any influential role; any reasonably small value of $\rho$ seems to yield similar results, and any $\tau^2$ which is not too close to 0 behaves similarly.

There seems to be strong evidence for a change-point in the intensity rate of casualties (induced by the ‘seat-belt law’) in the case of the LGV data, whereas this is absent in the case of the HGV data. This is evident from the very high value of $\log m_1$ near the ML-II estimate of $\delta^2$ for the LGV data.

There is a vast literature related to MDL, mostly in engineering and computer science. See [48] and the references listed there for the latest developments. See also [49]. [50] provides a review of MDL and SIC, and claim that SIC is the solution to “optimal universal coding problems”. MDL techniques have not become very popular in statistics, but they seem to be quite useful in many applications.

REFERENCES


Copyright © 2013 SciRes.


