Open Journal of Statistics, **2013, 3, 103-117**
http://dx.doi.org/10.4236/ojs.2013.32012 Published Online April 2013 (http://www.scirp.org/journal/ojs)

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

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 logaritm 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 δ

denotes the precision. In effect we will then be considering

$$P\left([x]-\delta/2 \le X \le [x]+\delta/2\right)=\int_{[x]-\delta/2}^{[x]+\delta/2} p(u)\,\mathrm{d}u \approx \delta p(x)$$

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 $p(\boldsymbol{x})$ by the probability of the r-dimensional cube of side δ containing $\boldsymbol{x}$, namely $p([\boldsymbol{x}])\delta^r \approx p(\boldsymbol{x})\delta^r$, so that the optimal code length changes to $-\log p(\boldsymbol{x}) - r\log\delta$.

3. MDL for Estimation or Model Fitting

Consider data $\boldsymbol{x} \equiv \boldsymbol{x}^n = (x_1, x_2, \cdots, x_n)$, and suppose

$$\mathscr{F} = \left\{ f\left(\boldsymbol{x}^n \middle| \theta\right) : \theta \in \Theta \right\}$$

is the collection of models of interest. Further, let $\pi(\theta)$ be a prior density for θ. Given a value of θ (or a model), the optimal code length for describing $\boldsymbol{x}^n$ is $-\log f\left(\boldsymbol{x}^n \middle| \theta\right)$, but since θ is unknown, its description requires a further $-\log \pi(\theta)$ bits on average. Therefore the optimal code length is obtained upon minimizing

$$DL(\theta) = -\log \pi(\theta) - \log f\left(\boldsymbol{x}^n \middle| \theta\right), \tag{1}$$

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 θ given the data $\boldsymbol{x}^n$ is

$$\pi\left(\theta \middle| \boldsymbol{x}^n\right) = \frac{f\left(\boldsymbol{x}^n \middle| \theta\right)\pi(\theta)}{m\left(\boldsymbol{x}^n\right)}, \tag{2}$$

where $m(\boldsymbol{y})$ is the marginal or predictive density. Therefore, minimizing

$$\begin{aligned} DL(\theta) &= -\log \pi(\theta) - \log f\left(\boldsymbol{x}^n \middle| \theta\right) \\ &= -\log\left\{ f\left(\boldsymbol{x}^n \middle| \theta\right)\pi(\theta)\right\} \end{aligned}$$

over θ is equivalent to maximizing $\pi\left(\theta \middle| \boldsymbol{x}^n\right)$. Thus MDL for estimation or model fitting is equivalent to finding the highest posterior density (HPD) estimate of θ. Note, however, that a prior π 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 $\Theta \subset \mathscr{R}^k$ for some fixed k (same k across all models), leading to the maximum likelihood estimate (MLE). The case of $\mathscr{F}$ 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^* and θ by $\boldsymbol{\theta}^k = (\theta_1, \theta_2, \cdots, \theta_k)$. Then

$$\begin{aligned} & DL^*\left(\boldsymbol{\theta}^k\right) \\ &= -\log\left\{\pi\left(\left[\boldsymbol{\theta}^k\right]_{\delta_\pi}\right)\delta_\pi^k\right\} - \log\left\{ f\left(\left[\boldsymbol{x}^n\right]_{\delta_f} \middle| \left[\boldsymbol{\theta}^k\right]_{\delta_\pi}\right)\delta_f^n\right\} \\ &= -\log \pi\left(\left[\boldsymbol{\theta}^k\right]_{\delta_\pi}\right) - k\log\delta_\pi - \log f\left(\left[\boldsymbol{x}^n\right]_{\delta_f} \middle| \left[\boldsymbol{\theta}^k\right]_{\delta_\pi}\right) \\ & -n\log\delta_f \\ &\approx -\log\pi\left(\boldsymbol{\theta}^k\right) - k\log\delta_\pi - \log f\left(\boldsymbol{x}^n \middle| \boldsymbol{\theta}^k\right) - n\log\delta_f. \end{aligned}$$

Here δ_f and δ_π are the precisions required to discretize x and θ, respectively. Note that the term $-n\log\delta_f$ is common across all models, so it can be ignored. However, the term $-k\log\delta_\pi$ which involves the dimension of θ in the model varies and is influential. According to [6,12], $\delta_\pi = 1/\sqrt{n}$ is optimal (see [13] for details), in which case

$$\begin{aligned} DL^*\left(\boldsymbol{\theta}^k\right) &\approx -\log f\left(\boldsymbol{x}^n \middle| \boldsymbol{\theta}^k\right) - \log\pi\left(\boldsymbol{\theta}^k\right) \\ &\quad + \frac{k}{2}\log n + \text{constant}. \end{aligned} \tag{3}$$

Minimizing this will not lead to MLE even when $\pi(\boldsymbol{\theta}^k)$ is assumed to be approximately constant. In fact, [12] proceeds further and argues that the correct precision δ_π should depend on the Fisher information matrix. This amounts to using a prior which is similar in nature to the Jeffreys' prior on θ. 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 θ, 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 and express it in the following form. Let $\boldsymbol{X} \equiv \boldsymbol{X}^n = (X_1, \cdots, X_n)$. Suppose $\boldsymbol{X}^n | \theta \sim f(\boldsymbol{x}^n | \theta), \theta \in \Theta$. Consider testing

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

where $\Theta_i \subset \Theta \subseteq \mathscr{R}^d$, for some $d, i = 0,1, \Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \phi$. Let π be a prior on Θ. Then π can be expressed as

$$\begin{aligned}\pi(\theta) &= \pi_0 g_0(\theta) I(\theta \in \Theta_0) \\ &\quad + (1-\pi_0) g_1(\theta) I(\theta \in \Theta_1)\end{aligned} \quad (5)$$

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

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

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

where

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

and

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

Note that m_i is simply the marginal or predictive density of $\boldsymbol{X}^n$ 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,

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

with BF_{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 BF_{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, $-\log m_i(\boldsymbol{x}^n)$, 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 $\boldsymbol{X}^n$ is a random sample from $N(\mu, \sigma^2)$ with known σ^2. We want to test

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

Consider the $N(0, \tau^2)$ prior on μ with known τ^2 under M_1. Then the marginal distribution of $\boldsymbol{X}^n$ is $N_n(\boldsymbol{0}, \sigma^2 I_n)$ under M_0 and under M_1 it is $N_n(\boldsymbol{0}, \sigma^2 I_n + \tau^2 \boldsymbol{11'})$. 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

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

where m_1 and m_0 are the corresponding densities. Since we are comparing two logarithms, let us switch to natural logarithms. Then

$$-\log m_0(\boldsymbol{x}^n) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2, \text{ and}$$

$$\begin{aligned}-\log m_1(\boldsymbol{x}^n) &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\left|\sigma^2 I_n + \tau^2 \boldsymbol{11'}\right| \\ &\quad + (\boldsymbol{x}^n)'(\sigma^2 I_n + \tau^2 \boldsymbol{11'})^{-1} \boldsymbol{x}^n.\end{aligned}$$

Noting that

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

and

$$\begin{aligned}(\sigma^2 I_n + \tau^2 \boldsymbol{11'})^{-1} &= \sigma^{-2}(I_n + (\tau^2/\sigma^2)\boldsymbol{11'})^{-1} \\ &= \sigma^{-2}(I_n - \tau^2/(\sigma^2 + n\tau^2)\boldsymbol{11'})\end{aligned}$$

we obtain

$$\begin{aligned}-\log m_1(\boldsymbol{x}^n) &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\left\{\sigma^{2n}(1 + n\tau^2/\sigma^2)\right\} \\ &\quad + \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 - \frac{n\tau^2}{2\sigma^2(n\tau^2 + \sigma^2)} n\bar{x}^2.\end{aligned}$$

Therefore M_1 is preferred over M_0 either by the Bayes factor or by the mixture MDL if and only if

$$\frac{n}{\sigma^2}\bar{x}^2 > \frac{n\tau^2+\sigma^2}{n\tau^2}\log\left(1+n\frac{\tau^2}{\sigma^2}\right).$$

Example 2. Let us consider the previous example with unknown σ^2 now. Suppose the prior on σ^2 is the default $1/\sigma^2$ under both models. The prior on μ under M_1 is now assumed to depend on σ^2, *i.e.*, $\mu|\sigma^2 \sim N\left(0,c\sigma^2\right)$, where $c>0$ is assumed to be a known constant for now. Then, provided $\boldsymbol{x}^n \neq \boldsymbol{0}$,

$$m_0\left(\boldsymbol{x}^n\right) = \int_0^\infty \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right)\frac{\mathrm{d}\sigma^2}{\sigma^2}$$
$$= \Gamma\left(\frac{n}{2}\right)\pi^{-n/2}\left(\sum_{i=1}^n x_i^2\right)^{-n/2},$$

and letting $S^2 = \sum_{i=1}^n\left(x_i-\bar{x}\right)^2$, and $m_1^*\left(\bar{x}|c,\sigma^2\right)$ denote the marginal density of $\bar{X}$ given c and σ^2,

$$m_1\left(\boldsymbol{x}^n\right)$$
$$= \int_0^\infty\int_{-\infty}^\infty \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{S^2}{2\sigma^2}\right)\left(2\pi c\sigma^2\right)^{-1/2}$$
$$\times\exp\left(-\left\{\frac{n\left(\mu-\bar{x}\right)^2}{2\sigma^2}+\frac{\mu^2}{2c\sigma^2}\right\}\right)\mathrm{d}\mu\frac{\mathrm{d}\sigma^2}{\sigma^2}$$
$$= \int_0^\infty\left(2\pi\sigma^2\right)^{-(n-1)/2}\exp\left(-\frac{S^2}{2\sigma^2}\right)m_1^*\left(\bar{x}|c,\sigma^2\right)\frac{\mathrm{d}\sigma^2}{\sigma^2}$$
$$= \Gamma\left(\frac{n}{2}\right)\pi^{-n/2}\left(1+nc\right)^{-1/2}\left(S^2+\frac{\bar{x}^2}{c+1/n}\right)^{-n/2}.$$

Thus

$$-\log\frac{m_1\left(\boldsymbol{x}^n\right)}{m_0\left(\boldsymbol{x}^n\right)}$$
$$= -\log\frac{\left(\sum_{i=1}^n x_i^2\right)^{-n/2}}{\left(S^2+\bar{x}^2/\left(c+1/n\right)\right)^{-n/2}\left(1+nc\right)^{-1/2}}$$
$$= -\frac{1}{2}\log\left(1+nc\right)+\frac{n}{2}\log\frac{S^2+n\bar{x}^2}{S^2+\bar{x}^2/\left(c+1/n\right)}$$
$$= -\frac{1}{2}\log\left(1+nc\right)+\frac{n}{2}\log\frac{1+n\bar{x}^2/S^2}{1+n\bar{x}^2/\left(\left(1+nc\right)S^2\right)}.$$

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 σ^2 is now replaced by an estimator.

Example 3. (**Jeffreys' Test**) This is similar to the problem discussed above, except that $\mu|\sigma \sim \mathscr{C}\left(0,\sigma\right)$, with density

$$g_1\left(\mu|\sigma\right) = \frac{1}{\sigma\pi}\frac{1}{\left(1+\mu^2/\sigma^2\right)},$$

the Cauchy prior. The prior on σ is the same as before under both models: $\pi_2\left(\sigma\right)=1/\sigma$. 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 $m_0\left(\boldsymbol{x}^n\right)$ is the same as in the previous example, namely

$$m_0\left(\boldsymbol{x}^n\right) = \int_0^\infty \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right)\frac{\mathrm{d}\sigma^2}{\sigma^2}$$
$$= \Gamma\left(\frac{n}{2}\right)\pi^{-n/2}\left(\sum_{i=1}^n x_i^2\right)^{-n/2},$$

whereas

$$m_1\left(\boldsymbol{x}^n\right)$$
$$= \int_0^\infty\int_{-\infty}^\infty\left(2\pi\sigma^2\right)^{-n/2}\mathrm{e}^{\left(-\frac{S^2}{2\sigma^2}-\frac{n\left(\mu-\bar{x}\right)^2}{2\sigma^2}\right)}\frac{1}{\sigma\pi\left(1+\mu^2/\sigma^2\right)}\mathrm{d}\mu\frac{\mathrm{d}\sigma}{\sigma}.$$

No closed form is available for $m_1(\boldsymbol{x}^n)$ in this case. To calculate this one can proceed as follows as indicated in Section 2.7 of [15]. The Cauchy density $g_1\left(\mu|\sigma\right)$ can be expressed as a Gamma scale mixture of normals,

$$g_1\left(\mu|\sigma\right) = \int_0^\infty\frac{\sigma}{\sqrt{2\pi}}\tau^{-1/2}\mathrm{e}^{-\sigma^2\tau/2}\left(\frac{\sqrt{\tau}}{\sqrt{2\pi}}\mathrm{e}^{-\tau\mu^2/2}\right)\mathrm{d}\tau$$

where τ is the mixing Gamma variable. Now one can integrate over μ and σ in closed form to simplify $(m_1\boldsymbol{x}^n)$. Finally, one has a one-dimensional integral over τ left, which can be numerically computed whenever needed.

Now, let us note from the examples discussed above that an efficient computation of $m_i(\boldsymbol{x}^n)$ 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}.

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 $m_i\left(\boldsymbol{x}^n\right)$ of equation (8). Let k_i be the dimension of Θ_i, $nh\left(\boldsymbol{\theta}^{k_i}\right) = -\log f\left(\boldsymbol{x}^n \middle| \boldsymbol{\theta}^{k_i}\right)$ and $\Delta_h\left(\boldsymbol{\theta}^{k_i}\right)$ denote the Hessian of h, *i.e.*,

$$\Delta_h\left(\boldsymbol{\theta}^{k_i}\right) = \left(-\frac{\partial^2}{\partial\theta_i\partial\theta_j} h\left(\boldsymbol{\theta}^{k_i}\right)\right)_{k_i\times k_i}.$$

Also, let $\hat{\boldsymbol{\theta}}^{k_i}$ denote either the MLE or the posterior mode. Then

$$\begin{aligned} &m_i\left(\boldsymbol{x}^n\right) \\ &= \int_{\Theta_i} f\left(\boldsymbol{x}^n \middle| \boldsymbol{\theta}^{k_i}\right) g_i\left(\boldsymbol{\theta}^{k_i}\right) \mathrm{d}\boldsymbol{\theta}^{k_i} \\ &= \mathrm{e}^{-nh\left(\hat{\boldsymbol{\theta}}^{k_i}\right)} (2\pi)^{k_i/2} n^{-k_i/2} \det\left(\Delta_h\left(\boldsymbol{\theta}^{k_i}\right)\right)^{-1/2} \\ &\times g_i\left(\boldsymbol{\theta}^{k_i}\right)\left(1+O\left(n^{-1}\right)\right), \end{aligned} \tag{11}$$

so that

$$\begin{aligned} &-\log m_i\left(\boldsymbol{x}^n\right) \\ &= nh\left(\hat{\boldsymbol{\theta}}^{k_i}\right) - \frac{k_i}{2}\log(2\pi) + \frac{k_i}{2}\log n \\ &+ \frac{1}{2}\log\det\left(\Delta_h\left(\hat{\boldsymbol{\theta}}^{k_i}\right)\right) - \log g_i\left(\hat{\boldsymbol{\theta}}^{k_i}\right) + O\left(n^{-1}\right) \\ &= -\log f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right) - \frac{k_i}{2}\log(2\pi) + \frac{k_i}{2}\log n \\ &+ \frac{1}{2}\log\det\left(\Delta_h\left(\hat{\boldsymbol{\theta}}^{k_i}\right)\right) - \log g_i\left(\hat{\boldsymbol{\theta}}^{k_i}\right) + O\left(n^{-1}\right). \end{aligned} \tag{12}$$

Ignoring terms that stay bounded as $n \to \infty$, [12] suggests using the (approximate) criterion which Rissanen calls *stochastic information complexity* (or "stochastic complexity" for short),

$$SIC\left(\boldsymbol{x}^n\right) = -\log f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right) + \frac{1}{2}\log\det\left(\hat{\Sigma}_n\right), \tag{13}$$

where

$$\hat{\Sigma}_n = n\Delta_h\left(\hat{\boldsymbol{\theta}}^{k_i}\right) \tag{14}$$

for implementing MDL. See [12,16-18] for further details.

If $X_1, X_2, \cdots, X_n$ are i.i.d. observations, then we have

$$\Delta_h\left(\hat{\boldsymbol{\theta}}^{k_i}\right) = I\left(\hat{\boldsymbol{\theta}}^{k_i}\right) + O\left(n^{-1}\right),$$

where $I(\boldsymbol{\theta})$ is the Fisher Information matrix and hence

$$\begin{aligned} -\log m_i\left(\boldsymbol{x}^n\right) &= -\log f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right) - \log g_i\left(\hat{\boldsymbol{\theta}}^{k_i}\right) \\ &+ \frac{k_i}{2}\log n + \frac{1}{2}\log\det\left(I\left(\hat{\boldsymbol{\theta}}^{k_i}\right)\right) \\ &- \frac{k_i}{2}\log(2\pi) + O\left(n^{-1}\right). \end{aligned} \tag{15}$$

Now ignoring terms that stay bounded as $n \to \infty$, we obtain the Schwarz criterion ([19]), BIC, for M_i given by

$$\mathrm{BIC}\left(\boldsymbol{x}^n\right) = \log f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right) - \frac{k_i}{2}\log n, \tag{16}$$

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:

$$M_0 : f\left(x \middle| \lambda, \alpha\right) = \lambda \exp\left(-\lambda x\right), x > 0$$

versus

$$M_1 : f\left(x \middle| \lambda, \alpha\right) = \lambda x^{\alpha-1} \exp\left(-\lambda x^\alpha\right), x > 0,$$

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

$$g_i\left(\boldsymbol{\theta}^{k_i}\right) = B_i\sqrt{\det\left(I\left(\boldsymbol{\theta}^{k_i}\right)\right)}, \boldsymbol{\theta}^{k_i} \in \Gamma_i,$$

where Γ_i is a compact subset of the relevant parameter space Θ_i and $B_i^{-1} = \int_{\Gamma_i}\sqrt{\det\left(I\left(\boldsymbol{\theta}^{k_i}\right)\right)}\mathrm{d}\boldsymbol{\theta}^{k_i}$. Consequently, it follows that,

$$\begin{aligned} m_i\left(\boldsymbol{x}^n\right) &= B_i \int_{\Gamma_i} f\left(\boldsymbol{x}^n \middle| \boldsymbol{\theta}^{k_i}\right)\sqrt{\det\left(I\left(\boldsymbol{\theta}^{k_i}\right)\right)}\mathrm{d}\boldsymbol{\theta}^{k_i} \\ &= B_i f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right)\left(\frac{n}{2\pi}\right)^{-k_i/2} + O\left(n^{-1}\right), \end{aligned}$$

where $\hat{\boldsymbol{\theta}}^{k_i}$ is the MLE of $\boldsymbol{\theta}^{k_i}$ under M_i. This yields,

$$\begin{aligned} -\log m_i\left(\boldsymbol{x}^n\right) &= -\log f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right) + \frac{k_i}{2}\log\frac{n}{2\pi} \\ &+ \log\int_{\Gamma_i}\left(\det\left(I\left(\boldsymbol{\theta}^{k_i}\right)\right)\right)^{1/2}\mathrm{d}\boldsymbol{\theta}^{k_i} + O\left(\frac{1}{n}\right). \end{aligned}$$

Compare this with (15) and note that the term involving $\det\left(I\left(\hat{\boldsymbol{\theta}}^{k_i}\right)\right)$ 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:

$$\frac{\sqrt{\det\left(I(\boldsymbol{\theta})\right)}}{\int_K\sqrt{\det\left(I(\boldsymbol{u})\right)}\mathrm{d}\boldsymbol{u}}, \text{for } \boldsymbol{\theta} \in K,$$

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 θ 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 $\hat{\boldsymbol{\theta}}^{k_i}$ be an estimate of $\boldsymbol{\theta}^{k_i}$ such as the posterior mean, HPD or MLE under M_i. This needs to be coded. Consider the prior density $g_i\left(\boldsymbol{\theta}^{k_i}\right)$ conditional on M_i being true. Usually MDL would choose a uniform density. Restrict $\boldsymbol{\theta}$ to a large compact subset of the parameter space and discretize it as discussed in Section 3 with a precision of $\delta_\pi = 1/\sqrt{n}$. Then the codelength required for coding $\hat{\boldsymbol{\theta}}^{k_i}$ is

$$L\left(\hat{\boldsymbol{\theta}}^{k_i}\right) = -\log g_i\left(\hat{\boldsymbol{\theta}}^{k_i}\right) + \frac{k_i}{2}\log n. \tag{17}$$

Stage 2. Now the data $\boldsymbol{x}^n$ is coded using the model density $f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right)$. Discretization may again be needed, say with precision δ_f. Thus the description length for coding $\boldsymbol{x}^n$ will be

$$-\log f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right) - n\log\delta_f. \tag{18}$$

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

$$\begin{aligned} &-\log f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right) - n\log\delta_f \\ &-\log g_i\left(\hat{\boldsymbol{\theta}}^{k_i}\right) + \frac{k_i}{2}\log n. \end{aligned} \tag{19}$$

Since the second term above, $-n\log\delta_f$, 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_i) as BIC, namely,

$$-\log f\left(\boldsymbol{x}^n \middle| \hat{\boldsymbol{\theta}}^{k_i}\right) + \frac{k_i}{2}\log n. \tag{20}$$

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 $\boldsymbol{y}^n$ on a response variable Y and also measurements $\boldsymbol{x}_1^n, \boldsymbol{x}_2^n, \cdots, \boldsymbol{x}_M^n$ on a set of potential explanatory variables (or regressors). Following [13], we associate with each regressor x_j, a binary variable γ_j. Then the set of available linear models is

$$\boldsymbol{y}^n = \sum_{\gamma_j:\gamma_j=1} \beta_j \boldsymbol{x}_j^n + \boldsymbol{\varepsilon}^n, \tag{21}$$

where $\boldsymbol{\varepsilon}^n \sim N_n\left(\boldsymbol{0}, \sigma^2 I_n\right)$. Note that $\boldsymbol{\gamma} = \left(\gamma_1, \gamma_2, \cdots, \gamma_M\right)$ is, then, a Bernoulli sequence associated with the set of regression coefficients, $\boldsymbol{\beta}^M = \left(\beta_1, \beta_2, \cdots, \beta_M\right)$ also. Let $\boldsymbol{\beta}_\gamma$ denote the vector of non-zero regression coefficients corresponding to $\boldsymbol{\gamma}$, and $\boldsymbol{X}_\gamma$ the corresponding design matrix, which results in the model

$$\boldsymbol{y}^n = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\varepsilon}^n.$$

Selecting the best model, then, is actually an estimation problem, *i.e.*, find the HPD estimate of $\boldsymbol{\gamma}$ starting with a prior π_1 on $\boldsymbol{\gamma}$ and a prior π_2 on $\left(\boldsymbol{\beta}_\gamma, \sigma^2\right)$ given $\boldsymbol{\gamma}$. The two-stage MDL, which is the simplest, uses the criterion of minimizing

$$\begin{aligned} DL\left(\boldsymbol{y}^n \middle| \boldsymbol{\gamma}\right) = &-\log f\left(\boldsymbol{y}^n \middle| \hat{\boldsymbol{\beta}}_\gamma, \hat{\sigma}_\gamma^2, \boldsymbol{X}_\gamma, \boldsymbol{\gamma}\right) \\ &-\log \pi_1\left(\boldsymbol{\gamma}\right). \end{aligned} \tag{22}$$

MLE for $\boldsymbol{\beta}_\gamma$ and σ^2 given $\boldsymbol{\gamma}$ are easily available:

$$\hat{\boldsymbol{\beta}}_\gamma = \left(\boldsymbol{X}'_\gamma \boldsymbol{X}_\gamma\right)^{-1} \boldsymbol{X}'_\gamma \boldsymbol{y},$$
$$\hat{\sigma}^2_\gamma = RSS_\gamma / n = \left\| \boldsymbol{y} - \boldsymbol{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma \right\|^2 / n. \quad (23)$$

Consider the uniform prior on $\boldsymbol{\gamma}$, 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)

$$\frac{n}{2}\log RSS_\gamma + \frac{k_\gamma}{2}\log n, \quad (24)$$

where k_γ is the number of $\{\gamma_j = 1\}$.

We can also derive the mixture MDL or stochastic complexity of a given model $\boldsymbol{\gamma}$. If $g\left(\boldsymbol{\beta}_\gamma, \sigma^2 \middle| \boldsymbol{\gamma}\right)$ is the prior density under π_2,

$$m\left(\boldsymbol{y}^n \middle| \boldsymbol{\gamma}\right) = \int f\left(\boldsymbol{y}^n \middle| \boldsymbol{\beta}_\gamma, \sigma^2, \boldsymbol{X}_\gamma, \boldsymbol{\gamma}\right) g\left(\boldsymbol{\beta}_\gamma, \sigma^2 \middle| \boldsymbol{\gamma}\right) \mathrm{d}\boldsymbol{\beta}_\gamma \mathrm{d}\sigma^2. \quad (25)$$

Applying (13), (14) after evaluating the information matrix of the parameters $\left(\boldsymbol{\beta}_\gamma, \sigma^2\right)$ and ignoring terms that are irrelevant for model selection, one obtains (see [13]),

$$\mathrm{SIC}(\boldsymbol{\gamma}) = \frac{n - k_\gamma - 2}{2}\log RSS_\gamma + \frac{k_\gamma}{2}\log n + \frac{1}{2}\log\det\left(\boldsymbol{X}'_\gamma \boldsymbol{X}_\gamma\right). \quad (26)$$

If g is chosen to be the conjugate prior density, then the marginal density $m\left(\boldsymbol{y}^n \middle| \boldsymbol{\gamma}\right)$ 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, $\boldsymbol{\gamma}$, one has to compute the corresponding $\hat{\boldsymbol{\beta}}_\gamma$ and RSS_γ.

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

$$M_0 : \boldsymbol{y}^n = \sum_{j=1}^{k} \beta_j \boldsymbol{x}_j^n + \boldsymbol{\varepsilon}^n$$

versus

$$M_1 : \boldsymbol{y}^n = \sum_{j=1}^{k+1} \beta_j \boldsymbol{x}_j^n + \boldsymbol{\varepsilon}^n.$$

This is actually a model building method, so we assume that $\boldsymbol{x}_1^n = \boldsymbol{1}^n$, and hence β_1 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 $x_1, \cdots, x_k$ 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

$$DL(k+1) - DL(k) \equiv \frac{n}{2}\log\left(\frac{RSS_{k+1}}{RSS_k}\right) + \frac{1}{2}\log n < 0, \quad (27)$$

where $DL(j)$ is the description length of the model with regressors $x_1, \cdots, x_j$ and 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.

$$\boldsymbol{X}_{(j)} = \left(\boldsymbol{x}_1^n \boldsymbol{x}_2^n \cdots \boldsymbol{x}_j^n\right), P^{(j)} = \boldsymbol{X}_{(j)}\left(\boldsymbol{X}'_{(j)}\boldsymbol{X}_{(j)}\right)^{-1}\boldsymbol{X}'_{(j)}.$$

Then we note the following result which may be found, for example, in [35].

$$\left(\boldsymbol{X}'_{(k+1)}\boldsymbol{X}_{(k+1)}\right)^{-1} = \begin{pmatrix} \boldsymbol{X}'_{(k)}\boldsymbol{X}_{(k)} & \boldsymbol{X}'_{(k)}\boldsymbol{x}_{k+1} \\ \boldsymbol{x}'_{k+1}\boldsymbol{X}_{(k)} & \boldsymbol{x}'_{k+1}\boldsymbol{x}_{k+1} \end{pmatrix}^{-1} = \begin{pmatrix} \left(\boldsymbol{X}'_{(k)}\boldsymbol{X}_{(k)}\right)^{-1} + v\boldsymbol{u}\boldsymbol{u}' & -v\boldsymbol{u} \\ -v\boldsymbol{u}' & v \end{pmatrix}, \quad (28)$$

where

$$\boldsymbol{u} = \left(\boldsymbol{X}'_{(k)}\boldsymbol{X}_{(k)}\right)^{-1}\boldsymbol{X}'_{(k)}\boldsymbol{x}_{k+1}$$
$$\text{and } v = \left(\boldsymbol{x}'_{k+1}\left(I_n - P^{(k)}\right)\boldsymbol{x}_{k+1}\right)^{-1}.$$

It then follows that

$$P^{(k+1)} = P^{(k)} + v\left(P^{(k)}\boldsymbol{x}_{k+1}\boldsymbol{x}'_{k+1}P^{(k)} - \boldsymbol{x}_{k+1}\boldsymbol{x}'_{k+1}P^{(k)} - P^{(k)}\boldsymbol{x}_{k+1}\boldsymbol{x}'_{k+1} + \boldsymbol{x}_{k+1}\boldsymbol{x}'_{k+1}\right)$$

and hence

$$RSS_k - RSS_{k+1} = \boldsymbol{y}'\left(P^{(k+1)} - P^{(k)}\right)\boldsymbol{y} = v\left\{\boldsymbol{y}'\left(I_n - P^{(k)}\right)\boldsymbol{x}_{k+1}\right\}^2,$$

and

$$\frac{RSS_k - RSS_{k+1}}{RSS_k} = \frac{\left(\boldsymbol{y}'\left(I_n - P^{(k)}\right)\boldsymbol{x}_{k+1}\right)^2}{\left\{\boldsymbol{x}'_{k+1}\left(I_n - P^{(k)}\right)\boldsymbol{x}_{k+1}\right\}\left\{\boldsymbol{y}'\left(I_n - P^{(k)}\right)\boldsymbol{y}\right\}} = r^2_{y, x_{k+1}\cdot(1,2,\cdots,k)},$$

where $r_{y,x_{k+1}\cdot(1,2,\cdots,k)}$ is simply the partial correlation coefficient between y and x_{k+1} conditional on $x_1,\cdots,x_k$. Substituting these in (27), we see that

$$\begin{aligned} &DL(k+1)-DL(k) \\ &=\frac{n}{2}\log\left(\frac{RSS_{k+1}}{RSS_k}\right)+\frac{1}{2}\log n \\ &=\frac{n}{2}\log\left(1-\frac{RSS_k-RSS_{k+1}}{RSS_k}\right)+\frac{1}{2}\log n \\ &=\frac{n}{2}\log\left(1-r^2_{y,x_{k+1}\cdot(1,2,\cdots,k)}\right)+\frac{1}{2}\log n. \end{aligned} \tag{29}$$

Therefore,

$DL(k+1)<DL(k)$ if and only if

$\frac{n}{2}\log\left(1-r^2_{y,x_{k+1}\cdot(1,2,\cdots,k)}\right)<-\frac{1}{2}\log n$ if and only if

$r^2_{y,x_{k+1}\cdot(1,2,\cdots,k)}>1-n^{-1/n}$.

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,

$$\begin{aligned} &\mathrm{SIC}(k+1)-\mathrm{SIC}(k) \\ &=\frac{n-k-2}{2}\log\left(\frac{RSS_{k+1}}{RSS_k}\right)-\frac{1}{2}\log RSS_{k+1} \\ &+\frac{1}{2}\log n+\frac{1}{2}\log\frac{\det\left(\boldsymbol{X}'_{(k+1)}\boldsymbol{X}_{(k+1)}\right)}{\det\left(\boldsymbol{X}'_{(k)}\boldsymbol{X}_{(k)}\right)} \\ &=\frac{n-k-2}{2}\log\left(1-r^2_{y,x_{k+1}\cdot(1,2,\cdots,k)}\right)+\frac{1}{2}\log n-\frac{1}{2}\log RSS_{k+1} \\ &+\frac{1}{2}\log\left\{\frac{\det\left(\boldsymbol{X}'_{(k)}\boldsymbol{X}_{(k)}\right)\left\{\boldsymbol{x}'_{k+1}\left(I_n-P^{(k)}\right)\boldsymbol{x}_{k+1}\right\}}{\det\left(\boldsymbol{X}'_{(k)}\boldsymbol{X}_{(k)}\right)}\right\} \\ &=\frac{n-k-2}{2}\log\left(1-r^2_{y,x_{k+1}\cdot(1,2,\cdots,k)}\right) \\ &+\frac{1}{2}\log n-\frac{1}{2}\log\left\{\frac{\boldsymbol{y}'\left(I_n-P^{(k+1)}\right)\boldsymbol{y}}{\boldsymbol{x}'_{k+1}\left(I_n-P^{(k)}\right)\boldsymbol{x}_{k+1}}\right\}, \end{aligned} \tag{31}$$

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

$$m_0\left(\boldsymbol{y}^n\right)=\int_0^\infty\int_{\mathscr{R}^k} f\left(\boldsymbol{y}^n\middle|\boldsymbol{\beta}^k,\boldsymbol{X}_{(k)},\sigma^2\right)g_0\left(\boldsymbol{\beta}^k,\sigma^2\right)\mathrm{d}\boldsymbol{\beta}^k\mathrm{d}\sigma^2,$$

and

$$\begin{aligned} &m_1\left(\boldsymbol{y}^n\right) \\ &=\int_0^\infty\int_{\mathscr{R}^{k+1}} f\left(\boldsymbol{y}^n\middle|\boldsymbol{\beta}^{k+1},\boldsymbol{X}_{(k+1)},\sigma^2\right)g_1\left(\boldsymbol{\beta}^{k+1},\sigma^2\right)\mathrm{d}\boldsymbol{\beta}^{k+1}\mathrm{d}\sigma^2 \end{aligned} \tag{32}$$

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,

$$\boldsymbol{\beta}^j\left|c,\sigma^2\sim N_j\left(\boldsymbol{0},c\sigma^2\left(\boldsymbol{X}'_j\boldsymbol{X}_j\right)^{-1}\right)\right.,\ j=k,k+1,$$

along with the density $1/\sigma^2$ for σ^2, a (proper prior) density $\pi(c)$ for the hyperparameter c, we obtain (30)

$$BF_{01}=\frac{\int_0^\infty m_0\left(\boldsymbol{y}^n\middle|c\right)\pi(c)\mathrm{d}c}{\int_0^\infty m_1\left(\boldsymbol{y}^n\middle|c\right)\pi(c)\mathrm{d}c}, \tag{33}$$

where

$$\begin{aligned} &m_0\left(\boldsymbol{y}^n\middle|c\right) \\ &=a_n(1+c)^{-k/2}\left[RSS_k+\frac{1}{1+c}\hat{\boldsymbol{\beta}}'_{(k)}\boldsymbol{X}'_{(k)}\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)}\right]^{-n/2}, \\ &m_1\left(\boldsymbol{y}^n\middle|c\right)=a_n(1+c)^{-(k+1)/2} \\ &\cdot\left[RSS_{k+1}+\frac{1}{1+c}\hat{\boldsymbol{\beta}}'_{(k+1)}\boldsymbol{X}'_{(k+1)}\boldsymbol{X}_{(k+1)}\hat{\boldsymbol{\beta}}_{(k+1)}\right]^{-n/2}, \end{aligned} \tag{34}$$

with $a_n=\pi^{-n/2}\Gamma(n/2)$ and $\hat{\boldsymbol{\beta}}_{(j)}=\left(\boldsymbol{X}'_{(j)}\boldsymbol{X}_{(j)}\right)^{-1}\boldsymbol{X}'_{(j)}\boldsymbol{y}$. The one-dimensional integrals in (33), however, cannot be obtained in closed form. One could also approximate BF_{01} with $m_0\left(\boldsymbol{y}^n\middle|\hat{c}_k\right)/m_1\left(\boldsymbol{y}^n\middle|\hat{c}_{k+1}\right)$, 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 = Year, 1 denoting 1930, X_2 = Pre-season precipitation, X_3 = May temperature, X_4 = June rain, X_5 = June temperature, X_6 = July rain, X_7 = July temperature, X_8 = August rain, X_9 = August temperature, and $Y = X_{10}$ = 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 $r^2_{y,j\cdot(1,\cdots,k)} > 1-n^{-1/n}$ (=0.1005 in this example). From this set we pick the one with the largest $r^2_{y,j\cdot(1,\cdots,k)}$. The values of $r^2_{y,j\cdot(1,\cdots,k)}$ for the relevant steps are listed below.

step	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	0.037	0.010	0.021	0.022	0.338	0.336	0.044	0.118
2	0.066	0.034	0.001	0.015	--	0.162	0.060	0.018
3	0.016	0.011	0.022	0.000	--	--	0.050	0.004

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

$$v(j,k) = r^2_{y,j\cdot(1,\cdots,k)} + \left\{\frac{\boldsymbol{y}'\left(I_n - P^{(k+1)}\right)\boldsymbol{y}}{n\boldsymbol{x}_j'\left(I_n - P^{(k)}\right)\boldsymbol{x}_j}\right\}^{-1/(n-k-2)} - 1,$$

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

step	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	−0.002	−0.020	0.043	−0.001	0.371	0.319	0.071	0.106
2	0.013	−0.011	0.012	−0.022	--	0.141	0.073	−0.002
3	−0.040	−0.040	0.032	−0.042	--	--	0.058	−0.022
4	−0.050	−0.042	0.027	−0.039	--	--	--	−0.023
5	−0.043	−0.027	--	−0.028	--	--	--	−0.025

According to SIC our order of selection is X_6, X_7, X_8, X_4.

5.2. Wavelet Thresholding

Consider the nonparametric regression problem where we have the following model for the noisy observations $\boldsymbol{v} = (v_1, v_2, \cdots, v_n)$:

$$v_i = s(x_i) + \varepsilon_i, i = 1,\cdots,n, \text{and } x_i \in \mathscr{T}, \tag{35}$$

where ε_i are i.i.d. $N(0,\sigma^2)$ errors with unknown error variance σ^2, and s is a function (or signal) defined on some interval $\mathscr{T} \subset \mathscr{R}^1$. Assuming s is a smooth function satisfying certain regularities (see [15,39,40]), we have the wavelet decomposition of s:

$$s(x) = \sum \beta_j \psi_j(x), \tag{36}$$

where ψ_j are the wavelet functions and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots)$ 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 $\boldsymbol{v}$, we get the estimated wavelet coefficients, $\boldsymbol{x}^n = W\boldsymbol{v}$. Consider now the equivalent model:

$$\boldsymbol{x}^n = \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim N_n\left(\boldsymbol{0}, \left(\sigma^2/n\right)I_n\right)$.

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

$$M_0 : \text{number of non-zero } \beta_j \textit{ is } k_0$$

versus

$$M_1 : \text{number of non-zero } \beta_j \textit{ is } k_1$$

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

The prior distribution on the non-zero β_j is assumed to be i.i.d. $N(0,\tau^2), 1 \le j \le k_i$ under $M_i, i = 0,1$.

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

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

Finally, we let $z_j, 1 \le j \le n$ be i.i.d. $N(0,\tau^2)$ (g be the corresponding joint density), and define the following structure under M_i.

$$\pi(\boldsymbol{\gamma}) = 1\Big/\binom{n}{k_i}, \text{for all } \boldsymbol{\gamma} : \sum_{j=1}^n \gamma_j = k_i; \tag{37}$$

$$\boldsymbol{\beta}|\boldsymbol{\gamma}, k_i : \beta_j = \gamma_j z_j, 1 \le j \le n. \tag{38}$$

$$\begin{aligned} &f\left(\boldsymbol{x}^n \middle| \boldsymbol{\beta}, \sigma^2\right) \\ &= f\left(\boldsymbol{x}^n \middle| \boldsymbol{z}, \boldsymbol{\gamma}, k_i, \sigma^2\right) \\ &= \left(2\pi\sigma^2/n\right)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}\sum_{j=1}^n \left(x_j - \gamma_j z_j\right)^2\right). \end{aligned} \tag{39}$$

The nuisance parameter σ^2 which is common under both models is given the prior density $1/\sigma^2$. Then it follows that

$$
\begin{aligned}
m_i\left(\boldsymbol{x}^n\right) &= m_i\left(\boldsymbol{x}^n \middle| k_i\right) \\
&= \sum\nolimits_{\gamma\in\Gamma_i} \frac{1}{\binom{n}{k_i}} \int_{\mathscr{R}^n\times\mathscr{R}^+} f\left(\boldsymbol{x}^n \middle| \boldsymbol{z},\boldsymbol{\gamma},k_i,\sigma^2\right) g\left(\boldsymbol{z}\right)\mathrm{d}\boldsymbol{z}\frac{\mathrm{d}\sigma^2}{\sigma^2} \\
&= \sum\nolimits_{\gamma\in\Gamma_i} \frac{1}{\binom{n}{k_i}} \int_{\mathscr{R}^n\times\mathscr{R}^+} \left(2\pi\sigma^2/n\right)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}\sum_{j=1}^{n}\left(x_j-\gamma_j z_j\right)^2\right)\times\left(2\pi\tau^2\right)^{-n/2}\exp\left(-\frac{1}{2\tau^2}\sum_{j=1}^{n} z_j^2\right)\mathrm{d}\boldsymbol{z}\frac{\mathrm{d}\sigma^2}{\sigma^2} \\
&= \sum\nolimits_{\gamma\in\Gamma_i} \frac{1}{\binom{n}{k_i}} \int_{\mathscr{R}^{k_i}\times\mathscr{R}^+} f\left(\boldsymbol{x}^n \middle| \boldsymbol{z},\boldsymbol{\gamma},k_i\right) g\left(\boldsymbol{z}\right)\mathrm{d}\boldsymbol{z}\frac{\mathrm{d}\sigma^2}{\sigma^2} \\
&= \sum\nolimits_{\gamma\in\Gamma_i} \frac{1}{\binom{n}{k_i}} \int_{\mathscr{R}^{k_i}\times\mathscr{R}^+} \left(2\pi\sigma^2/n\right)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}\sum_{j=1}^{n}\left(x_j-\gamma_j z_j\right)^2\right)\times\left(2\pi\tau^2\right)^{-k_i/2}\exp\left(-\frac{1}{2\tau^2}\sum_{j=1}^{n} \gamma_j z_j^2\right)\mathrm{d}\boldsymbol{z}\frac{\mathrm{d}\sigma^2}{\sigma^2}.
\end{aligned}
\tag{40}
$$

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 $m_i\left(\boldsymbol{x}^n\right)$ 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 β_j and σ^2 to a precision of $1/\sqrt{n}$ and ignoring terms that stay bounded as n increases, this amounts to

$$
\begin{aligned}
&DL\left(\boldsymbol{x}^n \middle| k_i\right) \\
&= -\log\pi\left(\hat{\boldsymbol{\gamma}}\middle| k_i\right) - \log\pi\left(\hat{\boldsymbol{\beta}}\middle|\hat{\boldsymbol{\gamma}},k_i\right) \\
&\quad -\log\pi\left(\hat{\sigma}^2\right) - \log f\left(\boldsymbol{x}^n\middle|\hat{\boldsymbol{\beta}},\hat{\sigma}^2\right) \\
&\approx \log\binom{n}{k_i} + \frac{1}{2\tau^2}\sum_{j=1}^{n}\hat{\beta}_j^2 \\
&\quad +\frac{n}{2}\log\left(2\pi\frac{\hat{\sigma}^2}{n}\right) + \frac{n}{2\hat{\sigma}^2}\sum_{j=1}^{n}\left(x_j-\hat{\beta}_j\right)^2 \\
&\approx k_i\log n + \frac{k_i}{2}\log n + \frac{n}{2}\log\left(2\pi\frac{\hat{\sigma}^2}{n}\right),
\end{aligned}
\tag{41}
$$

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 β_j's and $\hat{\sigma}^2$ is an estimate such as the MLE. On the other hand, computing SIC or $m_i\left(\boldsymbol{x}^n\right)$ is not an impossible task either. In fact, to integrate out the z_j in $m_i\left(\boldsymbol{x}^n\right)$ of Equation (40) we argue as follows.

$$X_j\left|\left(\gamma_j = 0,\sigma^2\right) \sim N\left(0,\sigma^2/n\right),\right. \text{ and}$$

$$X_j\left|\left(\gamma_j = 1,\sigma^2\right) \sim N\left(0,\tau^2+\sigma^2/n\right).\right.$$

Now, as we argued in Jeffreys test, we take $\tau^2 = c\sigma^2$, and integrate out σ^2 also. This leaves us with the following expression where we have a sum over $\boldsymbol{\gamma}$.

$$
\begin{aligned}
&m_i\left(\boldsymbol{x}^n\right) \\
&= \sum_{\gamma\in\Gamma_i}\frac{1}{\binom{n}{k_i}}\int_0^\infty \left(2\pi\sigma^2/n\right)^{-n/2}\left(1+nc\right)^{-k_i/2} \\
&\quad\times\exp\left(-\frac{n}{2\sigma^2}\left\{\sum_{j=1}^{n}\left(1-\gamma_j\right)x_j^2 + \frac{1}{1+nc}\sum_{j=1}^{n}\gamma_j x_j^2\right\}\right)\frac{\mathrm{d}\sigma^2}{\sigma^2} \\
&= \sum_{\gamma\in\Gamma_i}\frac{1}{\binom{n}{k_i}}\int_0^\infty \pi^{-n/2}\Gamma\left(n/2\right)\left(1+nc\right)^{-k_i/2} \\
&\quad\times\left\{\sum_{j=1}^{n}\left(1-\gamma_j\right)x_j^2 + \frac{1}{1+nc}\sum_{j=1}^{n}\gamma x_j^2\right\}^{-n/2}
\end{aligned}
\tag{42}
$$

The term $\left\{\sum_{j=1}^{n}\left(1-\gamma_j\right)x_j^2 + \frac{1}{1+nc}\sum_{j=1}^{n}\left(1-\gamma_j\right)x_j^2\right\}^{-n/2}$ is interesting. Most of the contribution to this sum is expected from $\sum_{j=1}^{n}\left(1-\hat{\gamma}_j\right)x_j^2$ with $\hat{\gamma}_j = 1$ corresponding to the largest k_i of the x_j, which yields the MLE of σ^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 $\boldsymbol{\gamma}$. As is clear, weighted average over the space Γ_i 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 approximation to a mixture MDL corresponding to a certain prior, a few questions related to this prior remain. First of all, the prior assumption that z_j 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 $Y_1, Y_2, \cdots, Y_n$ be independent Poisson counts with $Y_i \sim \text{Poisson}(\lambda_i)$. λ_i are *a priori* considered related, and a joint multivariate normal prior distribution on their logarithm is assumed. Specifically, let $\nu_i = \log(\lambda_i)$ be the i th element of $\boldsymbol{\nu}$ and suppose

$$\boldsymbol{\nu} | \boldsymbol{\mu} \sim N_n(\boldsymbol{\mu}, \delta^2 \Gamma),$$

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

$$M_1 : \mu_1 = \mu_2 = \cdots = \mu_{n_1} = \xi_0,$$
$$\mu_{n_1+1} = \mu_{n_1+2} = \cdots = \mu_n = \xi_1,$$

$$\Gamma = \begin{pmatrix} (1-\rho) I_{n_1} + \rho \mathbf{1}_{n_1} \mathbf{1}_{n_1}' & \mathbf{0} \\ \mathbf{0} & (1-\rho) I_{n_2} + \rho \mathbf{1}_{n_2} \mathbf{1}_{n_2}' \end{pmatrix}$$
$$= \Gamma_1,$$

where n_1 is the possible change-point. We further let ξ_0, ξ_1 be i.i.d. $N(0,\tau^2)$. Note that δ^2, ρ and τ^2 are hyperparameters.

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

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\nu}) &= \exp\left(-\sum_{i=1}^n \left\{ e^{\nu_i} - \nu_i y_i \right\}\right) \Big/ \prod_{i=1}^n y_i ! \\ &= \exp\left(-\sum_{i=1}^n e^{\nu_i - \log y_i} e^{\log y_i} + \sum_{i=1}^n y_i \left\{\nu_i - \log y_i\right\} + \sum_{i=1}^n y_i \log y_i\right) \\ &= \exp\left(\sum_{i=1}^n y_i \left\{\nu_i^* - e^{\nu_i^*}\right\}\right) \exp\left(\sum_{i=1}^n y_i \log y_i\right), \end{aligned} \tag{43}$$

where $\nu_i^* = \nu_i - \log y_i$. Expanding (43) about $\mathbf{0}$, its maximum, in Taylor series and ignoring higher order terms in ν_i^*, we obtain

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\nu}) &\approx \exp\left(\sum_{i=1}^n y_i \left\{\log y_i - 1\right\}\right) \\ &\cdot \exp\left(-\frac{1}{2} \sum_{i=1}^n y_i \left(\log y_i - \nu_i\right)^2\right). \end{aligned} \tag{44}$$

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

$$\boldsymbol{x} \equiv (x_1, \cdots, x_n) = (\log y_1, \cdots, \log y_n), \text{ and}$$
$$W \equiv \text{diag}(w_1, \cdots, w_n) = \text{diag}(1/y_1, \cdots, 1/y_n).$$

Thus hierarchical Bayesian analysis of multivariate normal linear models is applicable (see [15,36,47]). We note that the hyper-parameters ρ and τ^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,

$$\begin{aligned} m_0(\boldsymbol{y}^n | \rho, \tau^2) &= \int f(\boldsymbol{y}|\boldsymbol{\nu}) g_0(\boldsymbol{\nu}, \xi_0, \delta^2) \mathrm{d}\boldsymbol{\nu} \mathrm{d}\xi_0 \mathrm{d}\delta^2 \\ &\approx \exp\left(\sum_{i=1}^n y_i \left\{\log y_i - 1\right\}\right) (2\pi)^{n/2} \det(W)^{1/2} \times \int (2\pi)^{-\frac{n}{2}} |W|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\nu})' W^{-1} (\boldsymbol{x}-\boldsymbol{\nu})} g_0(\boldsymbol{\nu}, \xi_0, \delta^2) \mathrm{d}\boldsymbol{\nu} \mathrm{d}\xi_0 \mathrm{d}\delta^2 \\ &= \exp\left(\sum_{i=1}^n y_i \left\{\log y_i - 1\right\}\right) (2\pi)^{n/2} \det(W)^{1/2} \\ &\times \int (2\pi)^{-\frac{n}{2}} |W|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\nu})' W^{-1} (\boldsymbol{x}-\boldsymbol{\nu})} (2\pi)^{-\frac{n}{2}} \left|\delta^2 \left\{(1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n'\right\}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{\nu} - \xi_0 \mathbf{1})' \left(\delta^2 \left\{(1-\rho) I_n + \rho \mathbf{1}_n \mathbf{1}_n'\right\}\right)^{-1} (\boldsymbol{\nu} - \xi_0 \mathbf{1})} g_0(\xi_0, \delta^2) \mathrm{d}\boldsymbol{\nu} \mathrm{d}\xi_0 \mathrm{d}\delta^2. \end{aligned} \tag{45}$$

Now, from multivariate normal theory, observe that

$$\begin{aligned}&\int(2\pi)^{-\frac{n}{2}}|W|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{v})'W^{-1}(\boldsymbol{x}-\boldsymbol{v})}(2\pi)^{-\frac{n}{2}}\left|\delta^2\left\{(1-\rho)I_n+\rho\mathbf{1}_n\mathbf{1}_n'\right\}\right|^{-\frac{1}{2}}\\&\times\exp\left\{-\frac{1}{2}(\boldsymbol{v}-\xi_0\mathbf{1})'(\delta^2\left\{(1-\rho)I_n+\rho\mathbf{1}_n\mathbf{1}')\right\}^{-1}(\boldsymbol{v}-\xi_0\mathbf{1})\right\}\mathrm{d}\boldsymbol{v}\\&=(2\pi)^{-\frac{n}{2}}\left|W+\delta^2\left\{(1-\rho)I_n+\rho\mathbf{1}_n\mathbf{1}_n'\right\}\right|^{-\frac{1}{2}}\times\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\xi_0\mathbf{1})'\left(W+\delta^2\left\{(1-\rho)I_n+\rho\mathbf{1}_n\mathbf{1}_n'\right\}\right)^{-1}(\boldsymbol{x}-\xi_0\mathbf{1})\right\},\end{aligned} \tag{46}$$

and subsequently that, ξ_0 can be integrated out as in Example 2. Thus,

$$\begin{aligned}m_0\left(\boldsymbol{y}^n\middle|\rho,\tau^2\right)&\approx\exp\left(\sum_{i=1}^n y_i\{\log y_i-1\}\right)\times(2\pi)^{\frac{n}{2}}\det(W)^{\frac{1}{2}}\times(2\pi)^{-\frac{n}{2}}\\&\times\int\left|W+\delta^2\left\{(1-\rho)I_n+\rho\mathbf{1}_n\mathbf{1}_n'\right\}+\tau^2\mathbf{1}_n\mathbf{1}_n'\right|^{-\frac{1}{2}}\times\mathrm{e}^{-\frac{1}{2}\boldsymbol{x}'\left(W+\delta^2\left\{(1-\rho)I_n+\rho\mathbf{1}_n\mathbf{1}_n'\right\}+\tau^2\mathbf{1}_n\mathbf{1}_n'\right)^{-1}\boldsymbol{x}}g_0\left(\delta^2\right)\mathrm{d}\delta^2\\&=\int m_0\left(\boldsymbol{y}^n\middle|\delta^2,\rho,\tau^2\right)g_0\left(\delta^2\right)\mathrm{d}\delta^2.\end{aligned} \tag{47}$$

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 δ^2. Again, from (47), an ML-II likelihood for δ^2 is given by $\hat{\delta}_0^2$ which maximizes

$$\left|W+\delta^2\left\{(1-\rho)I_n+\rho\mathbf{1}_n\mathbf{1}_n'\right\}+\tau^2\mathbf{1}_n\mathbf{1}_n'\right|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}\boldsymbol{x}'\left(W+\delta^2\left\{(1-\rho)I_n+\rho\mathbf{1}_n\mathbf{1}_n'\right\}+\tau^2\mathbf{1}_n\mathbf{1}_n'\right)^{-1}\boldsymbol{x}}.$$

For fixed ρ and τ^2 this involves only examining a smooth function of a single variable, δ^2, which is a simple computational task.

We proceed exactly as above to derive $m_1\left(\boldsymbol{y}^n\middle|\rho,\tau^2\right)$ also. Partition the required vectors and matrices as follows.

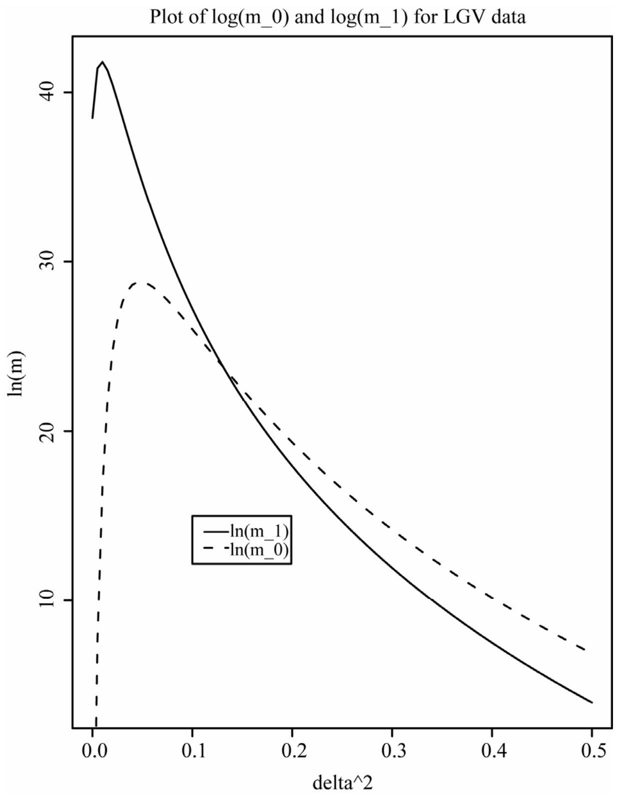
$$\boldsymbol{v}^n=\left(\boldsymbol{v}^{n_1},\boldsymbol{v}^{n_2}\right),\boldsymbol{y}^n=\left(\boldsymbol{y}^{n_1},\boldsymbol{y}^{n_2}\right),\boldsymbol{x}^n=\left(\boldsymbol{x}^{n_1},\boldsymbol{x}^{n_2}\right),$$

$$\boldsymbol{w}^n=\left(\boldsymbol{w}^{n_1},\boldsymbol{w}^{n_2}\right),W=\begin{pmatrix}W_{11}&\mathbf{0}\\\mathbf{0}&W_{22}\end{pmatrix},$$

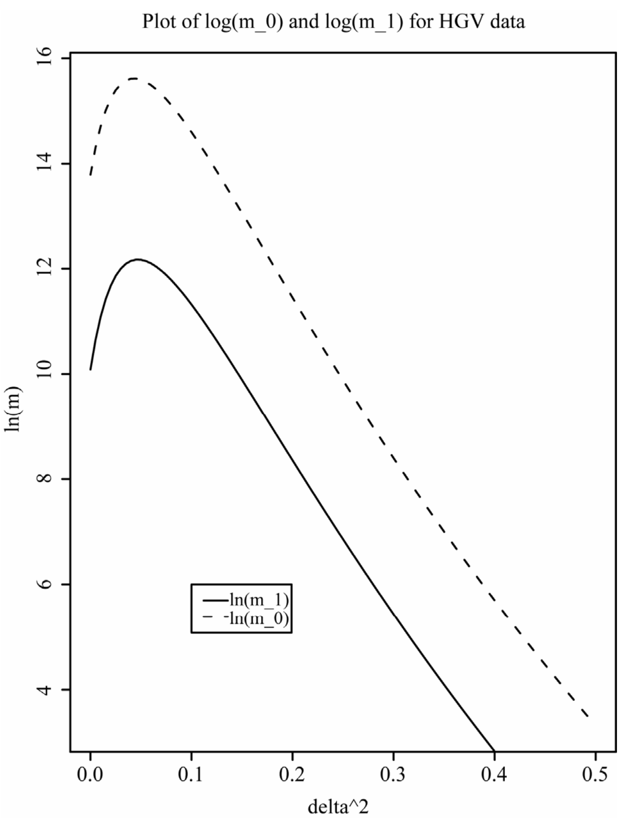
where $W_{11}=\operatorname{diag}\left(\boldsymbol{w}^{n_1}\right)$ and $W_{22}=\operatorname{diag}\left(\boldsymbol{w}^{n_2}\right)$.

Then we have

$$\begin{aligned}&m_1\left(\boldsymbol{y}^n\middle|\rho,\tau^2\right)=\int f\left(\boldsymbol{y}\middle|\boldsymbol{v}\right)g_1\left(\boldsymbol{v},\xi_0,\xi_1,\delta^2\right)\mathrm{d}\boldsymbol{v}\mathrm{d}\xi_0\mathrm{d}\xi_1\mathrm{d}\delta^2\\&\approx\exp\left(\sum_{i=1}^n y_i\{\log y_i-1\}\right)(2\pi)^{\frac{n}{2}}\det(W)^{\frac{1}{2}}\times\int(2\pi)^{-\frac{n}{2}}|W|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{v})'W^{-1}(\boldsymbol{x}-\boldsymbol{v})}g_1\left(\boldsymbol{v},\xi_0,\xi_1,\delta^2\right)\mathrm{d}\boldsymbol{v}\mathrm{d}\xi_0\mathrm{d}\xi_1\mathrm{d}\delta^2\\&=\exp\left(\sum_{i=1}^n y_i\{\log y_i-1\}\right)(2\pi)^{\frac{n}{2}}\det(W)^{\frac{1}{2}}\times\int(2\pi)^{-\frac{n_1+n_2}{2}}|W_{11}|^{-\frac{1}{2}}|W_{22}|^{-\frac{1}{2}}\times\exp\left\{-\frac{1}{2}\begin{pmatrix}\boldsymbol{x}^{n_1}-\boldsymbol{v}^{n_1}\\\boldsymbol{x}^{n_2}-\boldsymbol{v}^{n_2}\end{pmatrix}'\Gamma_1^{-1}\begin{pmatrix}\boldsymbol{x}^{n_1}-\boldsymbol{v}^{n_1}-\xi_0\mathbf{1}^{n_1}\\\boldsymbol{x}^{n_2}-\boldsymbol{v}^{n_2}-\xi_1\mathbf{1}^{n_2}\end{pmatrix}\right\}\\&\times(2\pi)^{-\frac{n_1+n_2}{2}}\left|\delta^2\left\{(1-\rho)I_{n_1}+\rho\mathbf{1}_{n_1}\mathbf{1}_{n_1}'\right\}\right|^{-\frac{1}{2}}\left|\delta^2\left\{(1-\rho)I_{n_2}+\rho\mathbf{1}_{n_2}\mathbf{1}_{n_2}'\right\}\right|^{-\frac{1}{2}}\\&\times\exp\left\{-\frac{1}{2\delta^2}\begin{pmatrix}\boldsymbol{v}^{n_1}-\xi_0\mathbf{1}^{n_1}\\\boldsymbol{v}^{n_2}-\xi_1\mathbf{1}^{n_2}\end{pmatrix}'\Gamma_1^{-1}\begin{pmatrix}\boldsymbol{v}^{n_1}-\xi_0\mathbf{1}^{n_1}\\\boldsymbol{v}^{n_2}-\xi_1\mathbf{1}^{n_2}\end{pmatrix}\right\}\times g_1\left(v,\xi_0,\xi_1,\delta^2\right)\mathrm{d}\boldsymbol{v}\mathrm{d}\xi_0\mathrm{d}\xi_1\mathrm{d}\delta^2\\&=\exp\left(\sum_{i=1}^n y_i\{\log y_i-1\}\right)(2\pi)^{\frac{n}{2}}\det(W)^{\frac{1}{2}}\times\int(2\pi)^{-\frac{n_1}{2}}\left|W_{11}+\delta^2\left\{(1-\rho)I_{n_1}+\rho\mathbf{1}_{n_1}\mathbf{1}_{n_1}'\right\}+\tau^2\mathbf{1}_{n_1}\mathbf{1}_{n_1}'\right|^{-\frac{1}{2}}\\&\times\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^{n_1}\right)'\left(W_{11}+\delta^2\left\{(1-\rho)I_{n_1}+\rho\mathbf{1}_{n_1}\mathbf{1}_{n_1}'\right\}+\tau^2\mathbf{1}_{n_1}\mathbf{1}_{n_1}'\right)^{-1}\boldsymbol{x}^{n_1}\right\}\times(2\pi)^{-\frac{n_2}{2}}\left|W_{22}+\delta^2\left\{(1-\rho)I_{n_2}+\rho\mathbf{1}_{n_2}\mathbf{1}_{n_2}'\right\}+\tau^2\mathbf{1}_{n_2}\mathbf{1}_{n_2}'\right|^{-\frac{1}{2}}\\&\times\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^{n_2}\right)'\left(W_{22}+\delta^2\left\{(1-\rho)I_{n_2}+\rho\mathbf{1}_{n_2}\mathbf{1}_{n_2}'\right\}+\tau^2\mathbf{1}_{n_2}\mathbf{1}_{n_2}'\right)^{-1}\boldsymbol{x}^{n_2}\right\}g_1\left(\delta^2\right)\mathrm{d}\delta^2\\&=\int m_1\left(\boldsymbol{y}^n\middle|\delta^2,\rho,\tau^2\right)g_1\left(\delta^2\right)\mathrm{d}\delta^2.\end{aligned} \tag{48}$$



(a)



(b)

Figure 1. Plots of log(m0) and log(m1) for the British road casualties data. (a) LGV data; (b) HGV data.

As before, the MDL technique invloves deriving the ML-II estimator of δ^2 from $m_1\left(\boldsymbol{y}^n \middle| \delta^2, \rho, \tau^2\right)$, for fixed ρ and τ^2. Obtaining $\hat{\delta}_1^2$ (for fixed ρ and τ^2) which maximizes $m_1\left(\boldsymbol{y}^n \middle| \delta^2, \rho, \tau^2\right)$ is very similar to that for $\hat{\delta}_0^2$.

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

[1] N. B. Asadi, T. H. Meng and W. H. Wong, "Reconfigurable Computing for Learning Bayesian Networks," *Proceedings of 16th International ACM/SIGDA Symposium on Field Programmable Gate Arrays*, Monterey, 24-26 February, 2008, pp. 203-211.

[2] D. M. Chickering, "Learning Bayesian Networks is NP-Complete," In: D. Fisher and H.-J. Lenz, Eds., *Learning from Data: AI and Statistics, V*, Springer, Berlin, Heidelberg, New York, 1996, pp. 121-130.

[3] H. Younes, M. Delampady, B. MacGibbon and O. Cherkaoui, "A Hierarchical Bayesian Approach to the Estimation of Monotone Hazard Rates in the Random Censorship Model," *Journal of Statistical Research*, Vol. 41, No. 2, 2007, pp. 35-42.

[4] Y. M. Shtarkov, "Universal Sequential Coding of Single Messages," *Problems of Information Transmission*, Vol. 23, No. 3, 1987, pp. 3-17.

[5] J. Rissanen, "Modeling by Shortest Data Description," *Automatica*, Vol. 14, No. 5, 1978, pp. 465-471. doi:10.1016/0005-1098(78)90005-5

[6] J. Rissanen, "A Universal Prior for Integers and Estimation by Minimum Description Length," *Annals of Statistics*, Vol. 11, No. 2, 1983, pp. 416-431. doi:10.1214/aos/1176346150

[7] C. S. Wallace and P. R. Freeman, "Estimation and Inference by Compact Coding (with Discussion)," *Journal of the Royal Statistical Society*, Vol. 49, No. 3, 1987, pp. 240-265.

[8] P. M. B. Vitanyi and M. Li, "Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity," *IEEE Transactions on Information Theory*, Vol. 46, No. 2, 2000, pp. 446-464. doi:10.1109/18.825807

[9] M. Li and P. Vitanyi, "An Introduction to Kolmogorov Complexity and Its Applications," 3rd Edition, Springer, Berlin, 2008. doi:10.1007/978-0-387-49820-1

[10] T. M. Cover and J. A. Thomas, "Elements of Information Theory," Wiley, Hoboken, 2006.

[11] G. H. Choe, "Computational Ergodic Theory," Springer, New York, 2005.

[12] J. Rissanen, "Stochastic Complexity and Statistical Inquiry," World Scientific, Singapore, 1989.

[13] B. Yu and M. H. Hansen, "Model Selection and the Principle of Minimum Description Length," *Journal of the American Statistical Association*, Vol. 96, No. 454, 2001, pp. 746-774. doi:10.1198/016214501753168398

[14] H. Jeffreys, "Theory of Probability," 3rd Edition, Oxford University Press, New York, 1961.

[15] J. K. Ghosh, M. Delampady and T. Samanta, "An Introduction to Bayesian Analysis: Theory and Methods," Springer, New York, 2006.

[16] J. Rissanen, "Stochastic Complexity and Modeling," *Annals of Statistics*, Vol. 14, No. 3, 1986, pp. 1080-1100. doi:10.1214/aos/1176350051

[17] J. Rissanen, "Stochastic Complexity (with Discussion)," *Journal of the Royal Statistical Society* (*Series B*), Vol. 49, No. 3, 1987, pp. 223-265.

[18] J. Rissanen, "Fisher Information and Stochastic Complexity," *IEEE Transactions on Information Theory*, Vol. 42, No. 1, 1996, pp. 48-54. doi:10.1109/18.481776

[19] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics*, Vol. 6, No. 2, 1978, pp. 461-464. doi:10.1214/aos/1176344136

[20] J. Rissanen and G. Shedler, "Failure-Time Prediction," *Journal of Statistical Planning and Inference*, Vol. 66, No. 2, 1998, pp. 193-210. doi:10.1016/S0378-3758(97)00083-9

[21] H. Matsuzoe, J. Takeuchi and S. Amari, "Equiaffine Structures on Statistical Manifolds and Bayesian Statistics," *Differential Geometry and Its Applications*, Vol. 24, No. 6, 2006, pp. 567-578. doi:10.1016/j.difgeo.2006.02.003

[22] J. Takeuchi, "Characterization of the Bayes Estimator and the MDL Estimator for Exponential Eamilies," *IEEE Transactions on Information Theory*, Vol. 43, No. 4, 1997, pp. 1165-1174. doi:10.1109/18.605579

[23] J. O. Berger and L. R. Pericchi, "The Intrinsic Bayes Factor for Linear Models (with Discussion)," In: J. M. Bernardo, *et al*., Eds., *Bayesian Statistics*, Oxford University Press, London, 1996, pp. 25-44.

[24] D. P. Foster and E. I. George, "The Risk Inflation Criterion for Multiple Regression," *Annals of Statistics*, Vol. 22, No. 4, 1994, pp. 1947-1975. doi:10.1214/aos/1176325766

[25] D. P. Foster and R. A. Stine, "Local Asymptotic Coding and the Minimum Description Length," *IEEE Transactions on Information Theory*, Vol. 45, No. 4, 1999, pp. 1289-1293. doi:10.1109/18.761287

[26] P. H. Garthwaite and J. M. Dickey, "Elicitation of Prior Distributions for Variable-selection Problems in Regression," *Annals of Statistics*, Vol. 20, No. 4, 1992, pp. 1697-1719. doi:10.1214/aos/1176348886

[27] P. H. Garthwaite and J. M. Dickey, "Quantifying and Using Expert Opinion for Variable-Selection Problems in Regression (with Discussion)," *Chemometrics and Intelligent Laboratory Systems*, Vol. 35, No. 1, 1996, pp. 1-43. doi:10.1016/S0169-7439(96)00035-4

[28] E. I. George and D. P. Foster, "Calibration and Empirical Bayes Variable Selection," *Biometrika*, Vol. 87, No. 4, 2000, pp. 731-747. doi:10.1093/biomet/87.4.731

[29] M. H. Hansen and B. Yu, "Bridging AIC and BIC: An MDL Model Selection Criterion," *Proceedings of the IT Workshop on Detection*, *Estimation*, *Classification and Imaging*, Santa Fe, 24-26 February 1999, p. 63.

[30] T. J. Mitchell and J. J. Beauchamp, "Bayesian Variable Selection in Linear Regression (with Discussion)," *Journal of the American Statistical Association*, Vol. 83, No. 404, 1988, pp. 1023-1036. doi:10.1080/01621459.1988.10478694

[31] A. F. M. Smith and D. J. Spiegelhalter, "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society*, (*Series B*), Vol. 42, No. 2, 1980, pp. 213-220.

[32] A. Zellner and A. Siow, "Posterior Odds Ratios for Selected Regression Hypotheses," In: J. M. Bernardo, *et al*., Eds., *Bayesian Statistics*, University Press, Valencia, 1980, pp. 585-603.

[33] A. Zellner, "Posterior Odds Ratios for Regression Hypotheses: General Considerations and Some Specific Results," In: A. Zellner, Ed., *Basic Issues in Econometrics*, University of Chicago Press, Chicago, 1984, pp. 275-305.

[34] A. Zellner, "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions," In: P. K. Goel and A. Zellner, Eds., *Basic Bayesian Inference and Decision Techniques*: *Essays in Honor of Bruno de Finetti*, Amsterdam, 1986, pp. 233-243.

[35] G. A. F. Seber, "A Matrix Handbook for Statisticians," Wiley, Hoboken, 2008, PMid: 19043372.

[36] J. O. Berger, "Statistical Decision Theory and Bayesian Analysis," 2nd Edition, Springer-Verlag, New York, 1985. doi:10.1007/978-1-4757-4286-2

[37] L. A. Shaw and D. D. Durast, "Measuring the Effects of Weather on Agricultural Output," ERS-72, US Department of Agriculture, Washington DC, 1962.

[38] L. M. Thompson, "Weather and Technology in the Production of Corn and Soybeans," CAED Report 17, Iowa State University, Iowa, 1963.

[39] A. Antoniadis, I. Gijbels and G. Grégoire, "Model Selection Using Wavelet Decomposition and Applications," *Biometrika*, Vol. 84, No. 4, 1997, pp. 751-763.

doi:10.1093/biomet/84.4.751

[40] J.-F. Angers and M. Delampady, “Bayesian Nonparametric Regression Using Wavelets,” *Sankhyā*: *The Indian Journal of Statistics*, Vol. 63, No. 3, 2001, pp. 287-308.

[41] M. S. Crouse, R. D. Nowak and R. J. Baraniuk, “Wavelet-Based Signal Processing Using Hidden Markov Models,” *IEEE Transactions on Signal Processing*, Vol. 46, No. 4, 1998, pp. 886-902. doi:10.1109/78.668544

[42] F. Abramovich and T. Sapatinas, “Bayesian Approach to Wavelet Decomposition and Shrinkage,” In: P. Müller and B. Vidakovic, Eds., *Bayesian Inference in Wavelet-Based Models*: *Lecture Notes in Statistics*, Vol. 141, Springer, New York, 1999, pp. 33-50.

[43] B. Vidakovic, “Nonlinear Wavelet Shrinkage with Bayes Rules and Bayes Factors,” *Journal of the American Statistical Association*, Vol. 93, No. 441, 1998, pp. 173-179. doi:10.1080/01621459.1998.10474099

[44] T. C. M. Lee, “Tree-Based Wavelet Regression for Correlated Data Using the Minimum Description Principle,” *Australian & New Zealand Journal of Statistics*, Vol. 44, No. 1, 2002, pp. 23-39. doi:10.1111/1467-842X.00205

[45] A. C. Harvey and J. Durbin, “The Effects of Seat Belt Legislation on British Road Casualties: A Case Study in Structural Time Series Modelling (with Discussion),” *Journal of the Royal Statistical Society* (*Series A*), Vol. 149, No. 3, 1986, pp. 187-227. doi:10.2307/2981553

[46] M. Delampady, I. Yee and J. V. Zidek, “Hierarchical Bayesian Analysis of a Discrete Time Series of Poisson Counts,” *Statistics and Computing*, Vol. 3, No. 1, 1993, pp. 7-15. doi:10.1007/BF00146948

[47] D. V. Lindley and A. F. M. Smith, “Bayes Estimates for the Linear Model,” *Journal of the Royal Statistical Society* (*Series B*), Vol. 34, 1972, pp. 1-41.

[48] P. D. Grunwald, “The Minimum Description Length Principle,” MIT Press, Cambridge, Massachusetts, 2007.

[49] S. de Rooij and P. Grunwald, “An Empirical Study of Minimum Description Length Model Selection with Infinite Parametric Complexity,” *Journal of Mathematical Psychology*, Vol. 50, No. 2, 2006, pp. 180-192. doi:10.1016/j.jmp.2005.11.008

[50] A. Barron, J. Rissanen and B. Yu, “The Minimum Description Length Principle in Coding and Modeling,” *IEEE Transactions on Information Theory*, Vol. 44, No. 6, 1998, pp. 2743-2760. doi:10.1109/18.720554

