The Consistency of MDL for Linear Regression Models with Increasing Signal-to-Noise Ratio

Daniel F. Schmidt, *Member*, *IEEE*, and Enes Makalic

***Abstract*—Recent work by Ding and Kay has demonstrated that the Bayesian information criterion (BIC) is an inconsistent estimator of model order in nested model selection as the noise variance $\tau^* \to 0$. Unfortunately, Ding and Kay have erroneously concluded that the minimum description length (MDL) principle also leads to inconsistent estimates of model order in this setting by equating BIC with MDL. This paper shows that only the earlier MDL criterion based on asymptotic assumptions has this problem, and proves that the new MDL linear regression criteria based on normalized maximum likelihood and Bayesian mixture codes satisfy the notion of consistency as $\tau^* \to 0$. The main result may be used as a basis to easily establish similar consistency results for other closely related information theoretic regression criteria.**

***Index Terms*—Model Selection, Consistency, Minimum Description Length, Linear Models**

I. Introduction

A model selection criterion is consistent if it almost surely selects the true generating model as the sample size $n \to \infty$, assuming the true data generating model is included in the set of candidate models under consideration. A recent paper by Ding and Kay [1] has examined a slightly different notion of model selection consistency in the context of linear regression models, in which the sample size n is fixed and the noise variance τ^* tends to zero. In the linear regression setting, the data $\mathbf{y} = (y_1, \ldots, y_n)'$ is assumed to have been generated from a linear combination of explanatory variables, or covariates, with additive noise, that is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_q)$ is an $(n \times q)$ design matrix, $\boldsymbol{\beta}^* \in \mathbb{R}^q$ is a vector of unknown, true, regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is a vector of disturbances, distributed as per $\varepsilon_i \sim N(0, \tau^*)$. It is common to assume that the model selection problem is nested, in the sense that the sub-design matrices considered by the model selection criteria are of the form $\mathbf{X}_k = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, for $k \in \{0, \ldots, q\}$, and that $\beta_i^* = 0$ for all $i > k^*$. The task of a model selection criterion is to infer the true order k^*.

A standard approach to estimating k^* is by minimisation of a penalized negative log-likelihood. These methods advocate choosing the model that minimises the sum of the negative log-likelihood and a suitable complexity penalty, that is,

$$\hat{k} = \underset{k \in \{0,\ldots,q\}}{\arg\min} \left\{ \frac{n}{2} \log \hat{\tau}_k + \alpha_k \right\},$$

where $\hat{\tau}_k$ is the maximum likelihood estimate of τ^* for model order k given by

$$\hat{\tau}_k = \frac{\mathbf{y}' \left(\mathbf{I}_n - \mathbf{X}_k (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k' \right) \mathbf{y}}{n},$$

and α_k is the penalty term. For example, $\alpha_k = k$ for the Akaike information criterion (AIC) [2] and $\alpha_k = (k/2) \log n$ for the Bayesian information criterion (BIC) [3]. A consistent model selection procedure is usually defined as one which satisfies

$$\mathbb{P}\left\{ \hat{k} = k^* \right\} \to 1 \text{ as } n \to \infty.$$

It is well known that the BIC is consistent in this sense if the number of candidate models under consideration is finite [4], while the AIC is not. The BIC is also known in the literature as the Schwarz information criterion (SIC), and incorrectly as the minimum description length (MDL) criterion. This latter name is the guise under which Ding and Kay investigated the BIC, and they proved that estimate of k^* obtained by minimising the "MDL" criterion does not satisfy the alternate notion of consistency given by

$$\mathbb{P}\left\{ \hat{k} = k^* \right\} \to 1 \text{ as } \tau^* \to 0, \tag{2}$$

where τ^* is the variance of the additive Gaussian noise, and the maximum order model considered $q < n$.

Unfortunately, the paper erroneously concludes that minimising the description length does not lead to a consistent estimator of k^*, in the sense of (2). The problem with this conclusion is that the "MDL" criterion that they examine is an early asymptotic approximation to the description length for a model [5]. The minimum description length principle [6], [7] defines a large range of so-called "universal models" on which model selection should ideally be based. The aim of this short note is twofold: (i) to show that minimising more exact measures of description length leads to model selection criteria that are consistent in the sense of (2), and (ii) to point out in the literature that MDL is *not* just the same as the BIC.

II. Consistency of MDL with decreasing noise variance

The MDL model selection criterion studied by Ding and Kay, which we shall henceforth refer to as "MDL78", was derived by making several assumptions about the behaviour of the maximum likelihood estimates of the model parameters, as well as the Fisher information matrix $\mathbf{J}_k$ associated with the model. For a linear regression model of order k, the Fisher information matrix for the k regression coefficients $\boldsymbol{\beta}$ is given by

$$\mathbf{J}_k(\tau) = \frac{\mathbf{X}_k' \mathbf{X}_k}{\tau}. \tag{3}$$

The MDL78 criterion makes the assumption that

$$\lim_{n \to \infty} \left\{ \frac{\mathbf{J}_k(\hat{\tau}_k)}{n} \right\} = \boldsymbol{\Sigma}_k,$$

where $\boldsymbol{\Sigma}_k$ is a $(k \times k)$ positive definite matrix. This assumption allows the determinant of the information matrix to be rewritten as

$$|\mathbf{J}_k(\hat{\tau}_k)| = n^k \cdot O(1),$$

where $O(1)$ is a term independent of n, and may be discarded with little effect on the criterion if n is large.

However, in the notion of consistency considered by Ding and Kay, the sample size n is fixed and $\tau^* \to 0$. In this case the information about the parameters, as measured by $|\mathbf{J}_k(\tau^*)|$, is an increasing function of $1/\tau^*$, and it is not unreasonable to expect a model selection criterion to be consistent as $\tau^* \to 0$. However, this violates one of the key assumptions under which the MDL78 was derived: namely, that the parameters of the true generating model are fixed while the sample size is increasing. Thus, it is unsurprising that the MDL78 criterion is inconsistent as $\tau^* \to 0$.

Since the introduction of the MDL principle in 1978, Rissanen and collaborators have continuously refined the notion of description length. This has lead to correspondingly refined model selection criteria, including new description length formulae for linear regression models. The criterion derived in [8] is based on the normalized maximum likelihood (NML) distribution, and is given by

$$\left(\frac{n-k}{2} \right) \log \hat{\tau}_k + \frac{k}{2} \log \left(\frac{\hat{R}_k}{n} \right) - \log \Gamma \left(\frac{n-k}{2} \right) - \log \Gamma \left(\frac{k}{2} \right), \tag{4}$$

Daniel F. Schmidt and Enes Makalic are with the Centre for MEGA Epidemiology, The University of Melbourne, Carlton, VIC 3053, Australia
Email: {dschmidt,emakalic}@unimelb.edu.au

where

$$\begin{aligned}\hat{R}_k &= \mathbf{y}'\mathbf{y} - n\hat{\tau}_k \\ &= \mathbf{y}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{y} \end{aligned} \quad (5)$$

is the fitted sum-of-squares. The criterion (4) is not based on asymptotic approximations, and we show that an estimate of k^* found by minimising (4) is consistent as $\tau^* \to 0$. In the following text we shall use the symbol "c" as a generic placeholder for *all* constant terms that do not depend on τ^*.

Theorem 1: Let $\hat{k} \in \{1, \ldots, q\}$ be an estimate of $k^* \in \{1, \ldots, q\}$ found by minimising a criterion of the form

$$I(k) = \left(\frac{n-k}{2}\right)\log\hat{\tau}_k + \frac{k}{2}\log\hat{R}_k + c_k, \quad (6)$$

where $q < n$, and c_k denotes terms that are constant with respect to $\hat{\tau}_k$, but may depend on k. Then, $\hat{k}$ is a consistent estimate of k^* as $\tau^* \to 0$.

Proof: To show that the estimate $\hat{k}$ will not overfit as $\tau^* \to 0$ we compute the probability that a criterion of the form (6) would prefer a model of order $(k^* + j)$ to the model of order k^*, i.e., $\mathbb{P}\{I(k^*) > I(k^* + j)\}$, which after some simplification is given by

$$\mathbb{P}\left\{n\log\frac{\hat{\tau}_{k^*}}{\hat{\tau}_{k^*+j}} + k^*\log\frac{\hat{R}_{k^*}}{\hat{\tau}_{k^*}} > (k^*+j)\log\frac{\hat{R}_{k^*+j}}{\hat{\tau}_{k^*+j}} + c\right\}. \quad (7)$$

Define the random variables $S_i = \hat{\tau}_i/\tau^*$; as $\hat{\tau}_i \sim (\tau^*/n)\chi^2_{n-i}$ for $i \geq k^*$, the variables S_i are independent of τ^* for $i \geq k^*$. After exponentiation, (7) may be rewritten as

$$\mathbb{P}\left\{c\left(\frac{S_{k^*}^{n-k^*}}{S_{k^*+j}^{n-k^*-j}}\right)\left(\frac{\hat{R}_{k^*}^{k^*}}{\hat{R}_{k^*+j}^{k^*+j}}\right) > \frac{1}{(\tau^*)^j}\right\}. \quad (8)$$

Denote the entire left hand side of (8) by the random variable Y, and let X denote the first ratio on the left-hand side of (8). From (5), and the properties of the least-squares estimates and quadratic forms of normal variates, we have

$$\begin{aligned}\mathrm{E}\left[\hat{R}_k\right] &\to c, \\ \mathrm{var}(\hat{R}_k) &\to 0,\end{aligned}$$

as $\tau^* \to 0$, for all $k > 0$. This implies that the second ratio on the left hand side of (8) converges in probability to a constant as $\tau^* \to 0$, and therefore $Y \xrightarrow{d} cX$ as $\tau^* \to 0$. The variable X is independent of τ^*, and the right-hand side of (8) is unbounded from above as $\tau^* \to 0$, implying that the probability of overfitting is vanishingly small as $\tau^* \to 0$.

To show that the estimate $\hat{k}$ will not underfit as $\tau^* \to 0$, we compute the probability that a criterion of the form (6) would prefer a model of order $(k^* - j)$, $0 < j < k^*$, to a model of order k^*. This is given by

$$\mathbb{P}\left\{n\log\frac{\hat{\tau}_{k^*}}{\hat{\tau}_{k^*-j}} + k^*\log\frac{\hat{R}_{k^*}}{\hat{\tau}_{k^*}} > (k^*-j)\log\frac{\hat{R}_{k^*-j}}{\hat{\tau}_{k^*-j}} + c\right\}. \quad (9)$$

After simplification, this may be written as

$$\mathbb{P}\left\{c\left(\frac{S_{k^*}^{n-k^*}}{\hat{\tau}_{k^*-j}^{n-k^*+j}}\right)\left(\frac{\hat{R}_{k^*}^{k^*}}{\hat{R}_{k^*-j}^{k^*-j}}\right) > \frac{1}{(\tau^*)^{n-k^*}}\right\}. \quad (10)$$

Let the random variable Z denote the left hand side of (10). By the properties of the least-squares estimates

$$\begin{aligned}\mathrm{E}\left[\hat{\tau}_k\right] &\to c, \\ \mathrm{var}(\hat{\tau}_k) &\to 0,\end{aligned}$$

as $\tau^* \to 0$, for all $k < k^*$. Therefore, $Z \xrightarrow{d} c\,S_{k^*}^{n-k^*}$ as $\tau^* \to 0$, where the random variable $S_{k^*}^{n-k^*}$ is independent of τ^*, and it follows that the probability of underfitting (10) is vanishly small as $\tau^* \to 0$. Thus,

$$\lim_{\tau^*\to 0}\left\{\mathbb{P}\left\{\hat{k} = k^*\right\}\right\} = 1$$

which establishes the consistency of $\hat{k}$ as $\tau^* \to 0$. □

It is clear that the MDL linear regression criterion given by (4) is of the form (6), and is thus consistent as $\tau^* \to 0$. Theorem 1 can also be used, with minor modifications, to prove the consistency of other information theoretic linear regression criteria, such as the minimum message length criteria discussed in [9]. As an example, the next subsection details a proof of the consistency of the g-MDL criterion as $\tau^* \to 0$.

A. Consistency of g-MDL

Theorem 1 can also be used to verify that the g-MDL criterion of Hansen and Yu [10], which is based on the Bayesian mixture form of MDL, is consistent as $\tau^* \to 0$. Let

$$\hat{F}_k = \frac{(n-k)\hat{R}_k}{nk\hat{\tau}_k}$$

denote the F statistic commonly used in hypothesis testing approaches to covariate selection. If $\hat{F}_k > 1$, the g-MDL description length, up to constants, for the k-th order model, when $k > 0$, is given by

$$\mathrm{gMDL}(k) = \left(\frac{n-k}{2}\right)\log\left(\frac{n\hat{\tau}_k}{n-k}\right) + \frac{k}{2}\log\left(\frac{\hat{R}_k}{k}\right) + \log n. \quad (11)$$

Otherwise, the null model with no covariates is considered to be preferable to the k-th order model, i.e., there is insufficient evidence in the data to estimate the k-th order model. The null model has a description length, up to constants, of

$$\mathrm{gMDL}(0) = \frac{n}{2}\log\left(\hat{\tau}_0\right) + \frac{1}{2}\log n, \quad (12)$$

where $\hat{\tau}_0 = \mathbf{y}'\mathbf{y}/n$. The following theorem establishes consistency of the g-MDL criterion as $\tau^* \to 0$.

Theorem 2: Let $\hat{k} \in \{0, \ldots, q\}$ be an estimate of $k^* \in \{1, \ldots, q\}$ found by minimising the g-MDL criterion, where $q < n$. Then, $\hat{k}$ is a consistent estimate of k^* as $\tau^* \to 0$.

Proof: As the criterion (11) is of the same form as (6), we know from Theorem 1 that

$$\mathbb{P}\left\{\mathrm{gMDL}(k^*) > \mathrm{gMDL}(k)\right\} \to 0 \text{ as } \tau^* \to 0$$

for all $k > 0$ and $k \neq k^*$. Thus, it remains only to establish that both

$$\mathbb{P}\{\hat{F}_{k^*} > 1\} \to 1, \quad (13)$$

and

$$\mathbb{P}\{\mathrm{gMDL}(k^*) > \mathrm{gMDL}(0)\} \to 0, \quad (14)$$

as $\tau^* \to 0$. Using the properties of least squares, and recalling the definition of the random variables $S_i = \hat{\tau}_i/\tau^*$, we may write the probability that $\hat{F}_{k^*} > 1$ as

$$\mathbb{P}\left\{\frac{c}{S_{k^*}} > \tau^*\right\}. \quad (15)$$

As the left hand side of the inequality (15) is independent of τ^*, the probability clearly goes to one as $\tau^* \to 0$, proving (13).

To prove (14) we examine the probability that the null model, with description length given by (12), would be erroneously preferred to the model of order k^*; this is given by

$$\mathbb{P}\left\{ n \log \frac{\hat{\tau}_{k^*}}{\hat{\tau}_0} + k^* \log \frac{\hat{R}_{k^*}}{\hat{\tau}_{k^*}} > c \right\}. \tag{16}$$

After exponentiation (16) may be rewritten as

$$\mathbb{P}\left\{ c\, S_{k^*}^{n-k^*} \left(\frac{\hat{R}_{k^*}^{k^*}}{\hat{\tau}_0^n} \right) > \frac{1}{(\tau^*)^{n-k^*}} \right\}. \tag{17}$$

Let the random variable W denote the left hand side of (17). From the properties of the generating model (1) we have

$$\begin{aligned} \mathrm{E}\left[\hat{\tau}_0\right] &\rightarrow c, \\ \mathrm{var}(\hat{\tau}_0) &\rightarrow 0, \end{aligned}$$

as $\tau^* \rightarrow 0$. Therefore, $W \stackrel{d}{\rightarrow} c\, S_{k^*}^{n-k^*}$ as $\tau^* \rightarrow 0$, where the random variable $S_{k^*}^{n-k^*}$ is independent of τ^*, and it follows that the probability of preferring the null model to the model of order k^* is vanishly small as $\tau^* \rightarrow 0$. □

REFERENCES

[1] Q. Ding and S. Kay, "Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 1959–1969, 2011.

[2] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, December 1974.

[3] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[4] D. M. A. Haughton, "On the choice of a model to fit data from an exponential family," *The Annals of Statistics*, vol. 16, no. 1, pp. 342–355, March 1988.

[5] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, September 1978.

[6] ——, *Information and Complexity in Statistical Modeling*, 1st ed., ser. Information Science and Statistics. Springer, 2007.

[7] P. D. Grünwald, *The Minimum Description Length Principle*, ser. Adaptive Communication and Machine Learning. The MIT Press, 2007.

[8] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[9] D. Schmidt and E. Makalic, "MML invariant linear regression," in *The 22nd Australasian Joint Conference on Artificial Intelligence*, Melbourne, Australia, 2009, pp. 312–321.

[10] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.