

A GENERALIZATION OF MDL FOR CHOOSING ADAPTATION MECHANISM AND DESIGN PARAMETERS IN IDENTIFICATION

Fredrik Gustafsson

Department of Electrical Engineering, Linköping University, S-581 83 Linköping

Abstract. The minimum description length (MDL) has for some decades been known to be an efficient tool for choosing model structure. We will in this contribution generalize MDL to adaptive algorithms in system and signal identification. The parameter vector in these problems can either be considered as piecewise constant using segmentation and change detection algorithms or as time-varying estimated by recursive identification algorithms. MDL is derived as a measure of code length needed to transmit or store a signal. With the generalization we can compute not only the best model structure for the signal, but also when it pays off to use recursive identification and transmit the parameter updates together with the residuals, or if it is better to segment the signal and transmit the change points and the parameters. The approach opens an auto-tuning possibility, where the design parameters in the recursive identification and change detection methods can be optimized with respect to the code length.

1. INTRODUCTION

In this contribution, we consider the problem of transmitting or storing a signal as efficiently as possible. By efficiently we mean by using as few binary digits as possible. As a tool for this, we can use a mathematical model of the signal, which is known to the receiver or the device that reads the stored information. In the following, we will refer only to the transmission problem. The point with the mathematical model is that we do not transmit the signal itself, but rather the residuals from the model whose size is of considerably smaller magnitude if the model is good and thus fewer bits are required for attaining the specified accuracy at the receiver. The prize we have to pay for this, is that the parameters in the model need to be transmitted as well. That is, we have a compromise between sending as small residuals as possible using as few parameters as possible. An implicit trade-off is the choice of how many decimals that are needed when transmitting the parameters. This last trade-off is however signal independent and can be optimized for each problem class leading to an *optimal code length*.

More specifically, the problem can be stated as choosing a regression vector (model structure) φ_t and parameter vectors θ_t (constant, piecewise constant or time-varying) in the signal model

$$y_t = \varphi_t^T \theta_t + e_t,$$

where e_t is the residual.

The problem classes we will examine are listed below:

- Time invariant models where different model structures can be compared.
- Time varying models where different model structures, recursive identification methods and their design parameters can be compared.
- Piecewise constant models where different model structures, change detection algorithms to find the change points and their design parameters can be compared.

Besides the residuals that always have to be transmitted, the first model requires a parameter vector to be transmitted. The second model transmits the time update of the parameter vector, whose size should increase as the time variations in the model increase. The third model requires the change points to be transmitted together with a parameter vector for each segment. Clearly, the use of too many change points should be penalized.

The first approach has been thoroughly examined by Rissanen, see for instance (Rissanen 1978, Rissanen 1982). He developed the Minimum Description Length (MDL) criterion, which is a direct measure of the number of bits that are needed to represent a given signal as a function of the number of parameters, the number of data and the size of the residuals. We will extend the MDL criterion for the latter two cases. The point with this contribution is that we get an answer to not only what the most appropriate model structure is, but also when it pays

¹ Submitted to SYSID'97

off to use recursive identification and change detection. Another point is that the design variables can be optimized automatically as well, which is important since this is often a difficult tuning issue.

We would like to point out the following:

- There is no assumption that there is a true system which has constant, time-varying or piecewise constant parameter, rather we are looking for an algorithm that is able to describe data as well as possible.
- The generalized MDL can be used for standard system identification problems, just like MDL is often used for choosing the model structure. For instance, by taking a typical realization of a time-varying system we get a suggestion on which recursive identification algorithm to apply and how the design parameters should be chosen.
- As will be shown, the result of minimizing the description length yields a stochastic optimality in the maximum likelihood meaning as well.

The paper is organized as follows. In section 2 the main points in the derivation of MDL is summarized. Section 3 derives the generalized MDL for the piecewise constant parameter case, while Section 4 generalizes MDL for time-varying parameters. The relation to the maximum likelihood method is pointed out in Section 5. Section 6 presents a simulation study and Section 7 concludes the paper.

2. TIME-INVARIANT PARAMETERS

The summary of MDL below essentially follows the introduction section of (Rissanen 1982). Assume the measured signal y_t is modelled in a parametric family with measurement noise σ^2 . Let $L(\varepsilon^N, \theta)$ denote the code length for the signal $y^N = (y_1, y_2, \dots, y_N)$ using a model with a d -dimensional parameter vector θ . Here $\varepsilon^N = (\varepsilon_1, \dots, \varepsilon_N)$ denotes the set of prediction errors from the model. In a linear regression framework, we have

$$\varepsilon_t = y_t - \varphi_t^T \theta$$

$$\theta = \left(\sum_{t=1}^N \varphi_t \varphi_t^T \right)^{-1} \sum_{t=1}^N \varphi_t y_t$$

but other model structures are of course possible.

Generally we have

$$L(\varepsilon^N, \theta) = -\log p(\varepsilon^N, \theta)$$

where $p(\varepsilon^N, \theta)$ is the joint distribution of data and the parameters. This expression is optimized over the precision of the value of θ , leading to that each element in the

parameter vector can be represented by an integer, say n . The code length of this integer can be expressed as $-\log(p(n))$ for a suitable choice of density function. Rissanen now proposes a non-informative prior for integers as

$$p(n) \sim 2^{\log^*(n)}$$

where

$$\log^*(n) = \log n + \log \log n + \log \log \log n + \dots$$

where the sum is terminated at the first negative term.

With this prior, the optimal code length can be written

$$L(\varepsilon^N, \theta) = -\log p(\varepsilon^N | \theta) + \log \|\theta\|_P^d + \log C(k) \quad (1)$$

where only the fastest growing penalty term is included. Here $C(k)$ is the volume of the unit sphere in R^k , and $\|\theta\|_P = \theta^T P^{-1} \theta$. For linear regressions with Gaussian noise we have

$$L(\varepsilon^N, \theta) = \frac{1}{\sigma^2} \sum_{t=1}^N \varepsilon_t^2 + \log \|\theta\|_P^d + d \log N \quad (2)$$

The most common reference to MDL only includes the first two terms, which are also scaled by $1/N$. However, to be able to compare different assumptions on parameter variations we keep the third term for later use.

3. PIECEWISE CONSTANT PARAMETERS

As a motivation for this approach consider the following example.

Example 1. The GSM standard for mobile telephony says that the signal is segmented in batches of 160 samples, in each segment an eighth order AR model is estimated and the parameter values (in fact non-linear transformation of reflection coefficients) and prediction errors (or rather a model of the prediction errors) are transmitted to the receiver.

This is an adequate coding since typical speech signals are short-time stationary. Note that the segmentation is fixed beforehand and known to the receiver in GSM.

We consider segmentation in a somewhat wider context, where also the time points defining the segmentation are kept as parameters. That is, the information needed to transmit comprises the residuals, the parameter vector in each segment and the change points. Related segmentation approaches are given in (Kitagawa and Akaike 1978, Djuric 1992), where the BIC criterion (Akaike 1977, Schwartz 1978) is used. Since BIC is the

same as MDL if only the fastest growing penalty term is included, the criteria they present will give almost identical result as the MDL.

If we consider the change points as fixed, the MDL theory immediately gives

$$L(\varepsilon^N, \theta^n) = -\log p(\varepsilon^N, \theta | k^n) = \frac{1}{\sigma^2} \sum_{t=1}^N \varepsilon^2(t) + dn \log N + \sum_{i=1}^n \log \|\theta_i\|_{P_i}^d$$

because with a given segmentation we are facing n independent coding problems. Note that the number of parameters are nd , so d in the MDL criterion is essentially replaced by nd . The last term is still negligible if $N \gg n$.

The remaining question is what the cost for coding the integers k^n is. One can argue that these integers are also parameters leading to the use of $n(d+1)$ in MDL as done in (Kitagawa and Akaike 1978). Or one can argue that code length of integers is negligible compared to the real-valued parameters, leading to MDL with kn parameters as used in (Djuric 1992). However, the description length of these integers is straightforward to compute. Bayes law gives that

$$p(\varepsilon^N, \theta, k^n) = p(\varepsilon^N, \theta | k^n) p(k^n).$$

The code length should thus be increased by $-\log(p(k^n))$. The most reasonable prior now is a flat one for each k . That is,

$$p(k^n) = \frac{1}{N(N-1) \cdots (N-n+1)} \approx \frac{1}{N^n}$$

where we have assumed that the number of data is much larger than the number of segments. This prior corresponds to the code length

$$L(k^n) = n \log N.$$

That is, the MDL penalty term should in fact be $n(k+1) \log(N)/N$.

$$L(\varepsilon^N, \theta^n, k^n) \approx \frac{1}{\sigma^2} \sum_{t=1}^N \varepsilon^2(t) + (d+1)n \log N \quad (3)$$

4. TIME-VARYING PARAMETERS

Here we consider adaptive algorithms that can be written as

$$\theta_{t+1} = \theta_t + \Delta\theta_t$$

For linear regression, the update can be written

$$\Delta\theta_t = \mu P_t \varphi_t \varepsilon_t$$

which comprises RLS, LMS and the Kalman filter as special cases.

As a first try, one can argue that the parameter update $\Delta\theta_t$ is a sequence of real numbers just like the residuals and that the MDL criterion should be

$$L(\varepsilon^N, \Delta\theta^N) = \sum_{t=1}^N \left(\frac{\varepsilon_t^2}{\sigma^2} + \|\Delta\theta_t\|_{P_\Delta}^d \right) \quad (4)$$

where P_Δ is the covariance matrix of the update $\Delta\theta_t$. This criterion exhibits the basic requirements of a penalty term linear in the number of parameters. Clearly, there is a tradeoff between making the residuals small (requiring large updates if the underlying dynamics are rapidly time-varying) and making the updates small.

5. ML INTERPRETATIONS OF MDL

As was shown in (Schwartz 1978, Veres 1991, Gustafsson and Hjalmarsson 1995), the MDL criterion (2) is asymptotically equivalent to the maximum likelihood estimate of the model structure. In (Gustafsson 1996) it is shown that MDL for segmentation (3) is asymptotically (in the segment lengths) equivalent to the ML estimate of the change points, which is a kind of model structure. An open question is if (4) corresponds to the ML estimate of, for instance, the forgetting parameter in RLS. In this sense, we can claim that MDL as we propose it corresponds to the ML estimate for choosing a *generalized model structure*.

These facts indicate that there is an alternate interpretation of MDL and that it might be applied for comparing generalized model structures for other purposes than coding.

6. SIMULATIONS

The signal in this section will be a first order AR model,

$$y(t) = -a_1(t)y(t-1) + e(t)$$

The noise variance is $Ee^2 = 1$ and $N = 200$ data are simulated. The AR parameter will be either constant, piecewise constant or slowly time-varying. The parameter is estimated by LS, RLS, LMS and SEGM, respectively. For each method and design parameter, the loss function and code length are evaluated on all but the 20 first data samples to avoid possible influence of transients and initialization.

The RLS and LMS algorithms are standard, and the design parameters are the forgetting factor λ and step size μ , respectively. The SEGM algorithm below is suggested in (Gustafsson 1996) as an approximation of the *maximum a posteriori* estimate of the change points in a piecewise constant model. The design parameter is the probability p for a change at each time instant, and a small value tends to give fewer change points than a large one. As alternatives, low-complexity change detection algorithms (Basseville and Nikiforov 1993) can be applied.

6.1 Time-invariant AR model

Consider first the case of a constant AR parameter $a = -0.5$. Figure 1 shows MDL and the loss function

$$V = \frac{1}{N} \sum_{t=1}^N \varepsilon_t^2$$

as a function of the design parameter and the parameter tracking where the true parameter value is indicated by a dotted line. Table 1 summarizes the optimal design parameters and code lengths for this particular example.

Note that the optimal design parameter in RLS corresponds to the LS solution and that the step size of LMS is very small (compared to the ones to follow). All methods have approximately the same code length, which is the logical result.

Method	Optimal par.	MDL	V
LS	—	1.023	0.997
RLS	$\lambda = 1$	1.016	1.016
LMS	$\mu = 0.002$	1.015	1.015
SEGM	$p = 0.3$	1.011	0.958

Table 1. Optimal code length and design parameters for RLS, LMS and SEGM, respectively

6.2 Abruptly changing AR model

Consider the piecewise constant AR parameter

$$a_1(t) = \begin{cases} -0.5 & \text{if } t \leq 100 \\ 0.5 & \text{if } t > 100 \end{cases}$$

Figure 2 shows MDL and the loss function V as a function of the design parameter and the parameter tracking where the true parameter value is indicated by a dotted line. Table 2 summarizes the optimal design parameters and code lengths for this particular example. The segmentation algorithm which is able to capture the time

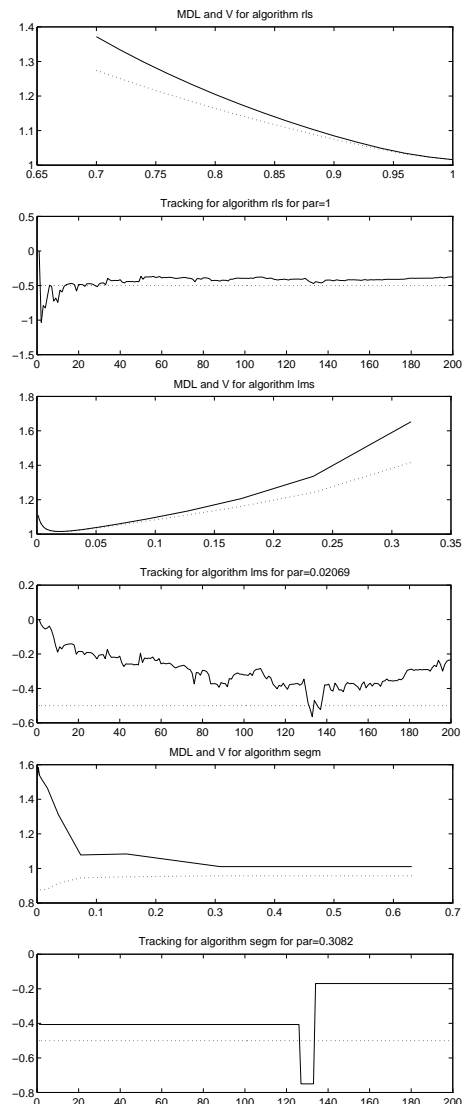


Fig. 1. MDL and V as a function of design parameter and parameter tracking for RLS, LMS and SEGM, respectively

variations perfectly comes out as the winner, but RLS and LMS are not much worse. Clearly, an adaptive algorithm is here much better than a fixed estimate. The updates $\Delta\theta(t)$ are of much smaller magnitude than the residuals.

Method	Optimal par.	MDL	V
LS	—	1.32	1.29
RLS	$\lambda = 0.94$	1.186	1.182
LMS	$\mu = 0.038$	1.189	1.186
SEGM	$p = 0.6$	1.138	1.085

Table 2. Optimal code length and design parameters for RLS, LMS and SEGM, respectively

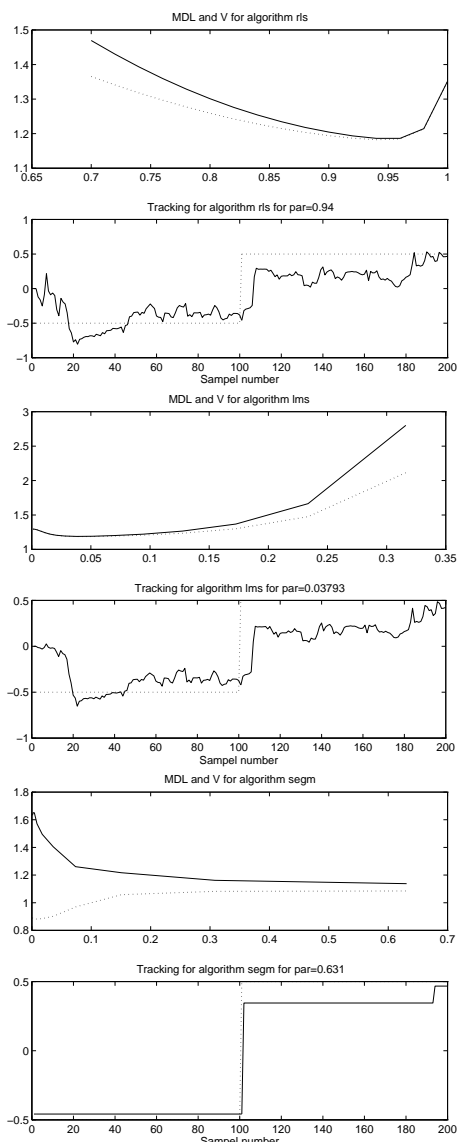


Fig. 2. MDL and V as a function of design parameter and parameter tracking for RLS, LMS and SEGM, respectively

6.3 Time-varying AR model

The simulation setup is exactly as before, but the parameter vector is linearly changing from -0.5 to 0.5 over 100 samples. Figure 3 and Table 3 summarize the result. This time the difference between the adaptive algorithms is less than before and there is no clear winner. The choice of adaptation mechanism is for this signal arbitrary.

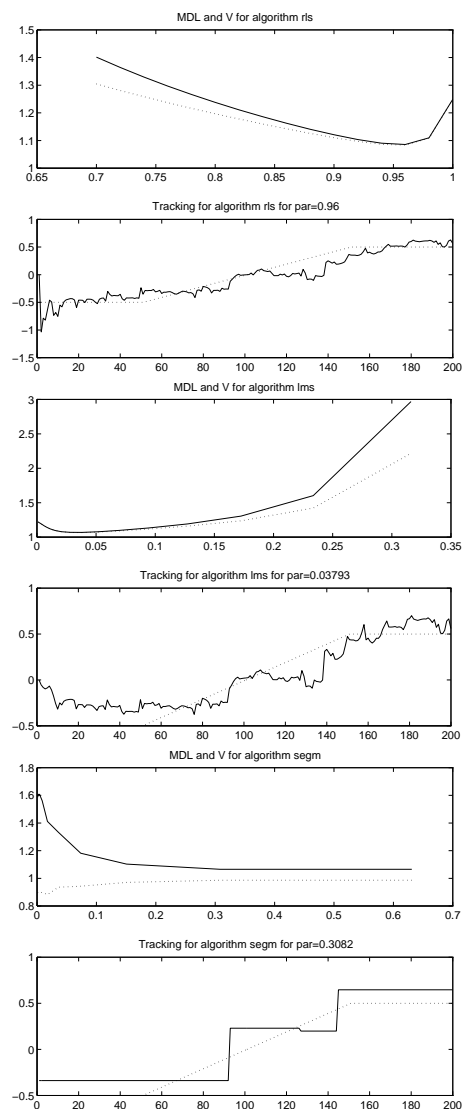


Fig. 3. MDL and V as a function of design parameter and parameter tracking for RLS, LMS and SEGM, respectively

Method	Optimal par.	MDL	V
LS	—	1.19	1.16
RLS	$\lambda = 0.96$	1.086	1.083
LMS	$\mu = 0.038$	1.069	1.066
SEGM	$p = 0.6$	1.066	0.986

Table 3. Optimal code length and design parameters for RLS, LMS and SEGM, respectively

7. CONCLUSIONS

We have studied a *generalized model structure* which not only includes the model structure parameterized in the parameter vector θ but also includes the adapta-

tion mechanism and its design parameters. This comprises algorithms where θ is considered as time-varying or piecewise constant. For the purpose of signal coding or storage, the code length is compared for these generalized model structures leading to a *generalized MDL* criterion. A simple simulation study showed the generalized MDL is capable to choose reasonable algorithms and design parameters for given signals.

8. REFERENCES

- Akaike, H. (1977). On entropy maximization principle. In: *Symposium on Applications of Statistics*.
- Basseville, M. and I.V. Nikiforov (1993). *Detection of abrupt changes: theory and application*. Information and system science series. Prentice Hall. Englewood Cliffs, NJ.
- Djuric, P. (1992). Segmentation of nonstationary signals. In: *Proceedings of ICASSP 92*. Vol. V. pp. 161–164.
- Gustafsson, F. (1996). Optimal segmentation of linear regression parameters. *Accepted for publication in IEEE Trans. on Signal Processing and available at <http://www.control.isy.liu.se/fredrik/journal>*.
- Gustafsson, F. and H. Hjalmarsson (1995). 21 ML estimators for model selection. *Automatica* **31**(10), 1377–1392.
- Kitagawa, G. and H. Akaike (1978). A procedure for the modeling of nonstationary time series. *Ann. Inst. Statist. Math.* **30**, 351–360.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14**, 465–471.
- Rissanen, J. (1982). Estimation of structure by minimum description length. *Circuits, Systems and Signal Processing* **1**, 395–406.
- Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
- Veres, S.M. (1991). *Structure Selection of Stochastic Dynamic Systems, the Information Criterion Approach*. Stochastic monographs. Gordon and Breach Science Publishers. New York.