A NEW INFORMATION THEORETIC APPROACH TO ORDER ESTIMATION PROBLEM

Soosan Beheshti [*] **Munther A. Dahleh** [*]

[*] *Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.*

Abstract: We introduce a new method of model order selection: minimum description complexity (MDC). The approach is motivated by the Kullback-Leibler information distance. The method suggests to choose the model set for which the "model set relative entropy" is minimum. The proposed method is comparable with the existing order estimation methods such as AIC and MDL. We elaborate on the advantages of MDC over the available information theoretic approaches.

Keywords: Kullback-Leibler distance, AIC, MDL, Order Estimation

1. INTRODUCTION

Classical problem of model selection among parametric model sets is considered. The goal is to choose a model set which best represents an observed data. The critical task is the choice of a criterion for model set comparison. Pioneer information theoretic based approaches to this problem are Akaike information criterion (AIC) and different forms of minimum description length (MDL) (Akaike, 1974), (Barron *et al.*, 1998). The prior assumption for calculation of these in criteria is that the unknown true model is a member of all the competing sets.

The new approach, minimum description complexity(MDC), is based on a new distance measure defined for the elements of the model sets. The distance of the true model and each model set is the minimum Kullback-Leibler distance of the true model and the elements of the model set. We provide a probabilistic method of MDC estimation for a class of parametric model sets. In this calculation the key factor is our prior assumption: unlike the existing methods no necessary assumption of the true parameter being a member of the competing model sets is needed. The main strength of the MDC calculation method is in its method of information extraction from the observed data.

Because of MDL's consistency, it has been widely used in practical problems. However, lack of a proper prior assumption in calculation of the criterion causes some defects such as high sensitivity to large signal to noise ratio when the true model does not belong to any of the model sets (A.P. Liavas and J.Delmas, 1999). Here we compare MDC with MDL and AIC in application. MDC is able to answer the challenging question of quality evaluation in identification of stable LTI systems under a fair prior assumption on the unmodeled dynamics. It also provides a new solution to a class of signal denoising problems (Beheshti and Dahleh, 2002).

2. IMPORTANT INFORMATION THEORETIC CRITERIA

Consider the following problem: Given a finite set of observed data y^N of length N, which is an element of set Y^N, a family of models which are parameterized by elements of a compact set S_M with order M, and a family of probability density functions $f(Y^N; S_M)$, select the model that best fits the data (Wax and Kailath, 1985).

In the following order estimation methods the estimate of the true parameter θ^* is calculated in S_m, a subset of S_M of order m. The estimate, $\hat{\theta}_{S_m}(y^N)$, is the maximum likelihood (ML) estimate of θ^* in S_m.

Akaike information criterion(AIC) is the estimate of the Kullback-Leibler distance of the true density

$f(Y^N;\theta)$, and the estimated density $f(Y^N;\hat{\theta}_{S_m}(y^N))$ in S_m (Akaike, 1974).

The AIC estimate is given by

$$\mathrm{AIC}_{S_m}(y^N) = -\frac{1}{N}\log f(y^N;\hat{\theta}_{S_m}(y^N)) + \frac{m}{N} \qquad (1)$$

This estimate is calculated with the assumption that N is large enough and that the parameter estimate, $\hat{\theta}_{S_m}(y^N)$, approaches the true parameter θ^* in subset S_m. The method suggests to select the model set S_m which minimizes the AIC.

Any model defined by a parameter in set S_M can be used to encode the observed data by using the Shannon coding method. The two-stage minimum description length(MDL) method is defined based on this coding. In the two-stage MDL approach the description length of the data in each subset is defined as (Rissanen, 1984)

$$\mathrm{DL}_{S_m}(y^N) = -\frac{1}{N}\log f(y^N;\hat{\theta}_{S_m}(y^N)) + m\frac{\log N}{2N}. \quad (2)$$

Similar to AIC the main assumption in calculation of MDL is that the ML estimate $\hat{\theta}_{S_m}(y^N)$ approaches θ^* as the length of data grows.

Bayesian information criterion(BIC) is another order estimation method which was proposed in (Schwarz, 1978). In this method a prior probability for the competing model sets is assumed. It is suggested to select the model that yields the maximum posterior probability. Note that the criterion in this approach is similar to MDL criterion in (2).

The two important prior assumptions for calculation of AIC and MDL in (1) and (2), for subset S_m, are

$$1)\ \theta^* \in S_m, \quad 2)\ \hat{\theta}_{S_m}(y^N) \to \theta^* \qquad (3)$$

The second condition in most cases implies that $M << N$.

Note that in practical problems we do not know whether or not the unknown θ^* is an element of a given S_m. However, in application of MDL and AIC the calculated criteria in (1) and (2) are used for all the subsets regardless of validity of the two prior assumptions in (3).

3. MINIMUM DESCRIPTION COMPLEXITY

We introduce a new method of subset selection by using the observed data of length N. Unlike the existing approaches none of the conditions in (3) are needed as our prior assumption. The set Y^N need not to be stationary. However, for each parametric probability distribution function (pdf), the expected value of Y^N and its covariance are finite. Also the pdfs $f(y^N;\theta)$ are continuous functions of Y^N.

Before providing a method of order estimation using the observed data we define a notion of distance for the pdfs. Note that in the following discussion the length of data is assumed to be fixed N.

For a given compact set S_M, use a positive cost function $V(\theta,y^N)$ for which $E_{\theta_1}\frac{1}{N}V(\theta_2,Y^N)$ is a finite nonnegative number and

$$E_{\theta_1}\frac{1}{N}V(\theta_2,Y^N) \geq E_{\theta_1}\frac{1}{N}V(\theta_1,Y^N) \qquad (4)$$

for any θ_1 and θ_2 in S_M. The equality holds only for $\theta_1 = \theta_2$. Choose the cost function such that it is a continuous function of both θ and y^N.

Definition 1 The description complexity of Y^N using pdf $f(Y^N;\theta_1)$, when the data is generated by θ, is defined by

$$\mathrm{DC}_N(\theta,\theta_1) \triangleq E_\theta \frac{1}{N}V(\theta_1,Y^N) \qquad (5)$$

For any element of S_M, define $\bar{\theta}_{S_m}$ in set S_m as

$$\bar{\theta}_{S_m}(N) = \arg\min_{\theta_1\in S_m} E_\theta \frac{1}{N}V(\theta_1,Y^N). \qquad (6)$$

The description complexity of the data set using subset S_m, when the data is generated by θ, is then defined as

$$\mathrm{DC}_N(\theta,S_m) \triangleq \min_{\theta_{S_m}\in S_m} \mathrm{DC}_N(\theta,\theta_{S_m}) \qquad (7)$$

$$= \mathrm{DC}_N\left(\theta,\bar{\theta}_{S_m}(N)\right) \qquad (8)$$

Definition 2 The minimum description complexity (MDC) of Y^N, when the data is generated by θ, is provided by subset S_m^*

$$S_m^* = \arg\min_{S_m} \mathrm{DC}_N\left(\theta,\bar{\theta}_{S_m}(N)\right). \qquad (9)$$

In general the set of all possible cost functions depends on the structure of the parametric model set. One example of such cost function for any parametric pdf is

$$V(\theta,y^N) = -\log f(y^N;\theta) \qquad (10)$$

This function is well defined for all Y^N with prior assumption that $f(y^N;\theta) \neq 0$ for any y^N. For this cost function

$$\mathrm{DC}_N(\theta,\theta) = E_\theta\frac{1}{N}V(\theta,Y^N) = \frac{1}{N}H_\theta(Y^N)\log 2 \qquad (11)$$

where $H_\theta(Y^N)$ is the differential entropy of Y^N when it is generated by θ

$$H_\theta(Y^N) = -E_\theta \log_2 f(y^N;\theta) \qquad (12)$$

If we want the description complexity function to be more like a distance measure we add the extra condition $\mathrm{DC}(\theta,\theta) = 0$. For example the new description complexity using the defined DC in (5) can be defined as

$$I_N(f(:;\theta),f(:,\bar{\theta}_{S_m}(N)) = \\ \mathrm{DC}_N(\theta,\bar{\theta}_{S_m}(N)) - \mathrm{DC}_N(\theta,\theta) \tag{13}$$

where the cost function is defined in (10). In this case $I_N(\cdot)$ is the Kullback-Leibler distance of θ and $\bar{\theta}_{S_m}(N)$.

3.1 *MDC and Data Observation*

Based on the defined description complexity, consider a family of parameter estimators for which

$$E_\theta(\hat{\theta}_{S_m}(Y^N)) = \bar{\theta}_{S_m}(N) \tag{14}$$

where $\bar{\theta}_{S_m}(N)$ is defined in (6). Note that for this set of estimators we have

$$E_\theta(\hat{\theta}_{S_M}(Y^N)) = \theta \tag{15}$$

and therefore the estimator is unbiased in S_M. The observed data y^N is generated by the unknown parameter θ^* and in each subset $\hat{\theta}_{S_m}(y^N)$ and $V(\hat{\theta}_{S_m}(y^N),y^N)$ are available. For order estimation the goal is to first use this information to find an estimate for $\mathrm{DC}(\theta^*,\hat{\theta}_{S_m}(y^N))$ for each subset and then choose the subset for which this error is minimum.

The first step is to validate θ's given the available estimate $\theta_{S_m}(y^N)$. The random variable $\mathrm{DC}(\theta,\hat{\theta}_{S_m}(Y^N))$ for each θ has a mean and a variance which are functions of θ, S_m and N. If the data is generated with θ then with probability p, the bound $\varepsilon_p(\theta,S_m,N)$ is available such that for a set of $x^N \in Y^N$

$$\Pr\{|(\mathrm{DC}(\theta,\hat{\theta}_{S_m}(x^N)) - \mathrm{E}_\theta \mathrm{DC}(\theta,\hat{\theta}_{S_m}(Y^N))| \\ \leq \varepsilon_p(\theta,S_m,N)\} = p \tag{16}$$

and subset $T_p(\theta,S_m,N)$ in Y^N is defined as

$$T_p(\theta,S_m,N) = \{x^N \in Y^N : \\ |(\mathrm{DC}(\theta,\hat{\theta}_{S_m}(x^N)) - \mathrm{E}_\theta \mathrm{DC}(\theta,\hat{\theta}_{S_m}(Y^N))| \\ \leq \varepsilon_p(\theta,S_m,N)\}.$$

The validation with probability p, and based on the observed data, provides the following set of parameters in S_M

$$\Theta(y^N,S_m,p) = \{\theta \in \Theta | y^N \in T_p(\theta,S_m,N)\}. \tag{17}$$

Therefore, with validation probability p in each subset S_m, the desired DC, $\mathrm{DC}_N(\theta^*,\hat{\theta}_{S_m}(y^N))$, is bounded by

$$\min_{\theta\in\Theta(y^N,S_m,p)} \mathrm{DC}_N(\theta,\hat{\theta}_{S_m}(y^N)) \leq \tag{18} \\ \mathrm{DC}_N(\theta^*,\hat{\theta}_{S_m}(y^N)) \leq \max_{\theta\in\Theta(y^N,S_m,p)} \mathrm{DC}_N(\theta,\hat{\theta}_{S_m}(y^N))$$

Note that if the observed y^N can be produced by all elements of S_M, then for $p=1$,

$$T_1(\theta,N,S_m) = S_M \tag{19}$$

and we have

$$0 \leq \mathrm{DC}_N(\theta^*,\hat{\theta}_{S_m}(y^N)) \leq \max_{\theta\in S_M} \mathrm{DC}_N(\theta,\hat{\theta}_{S_m}(y^N)) \tag{20}$$

However, if $p \neq 1$ the value of $\varepsilon_p(\theta,S_m,N)$ and therefore the set $T_p(\theta,N,S_m)$ depends on the *variance* of random variable $\mathrm{DC}(\theta,\hat{\theta}_{S_m}(Y^N))$. In most cases the variance of this error is a function of dimension (order) of S_m. With a fixed data of length N as the dimension m grows, the variance of error also grows. However the estimate bias is a decreasing function of order. Therefore for a given finite length data there is a tradeoff between the error variance and bias.

The MDC order estimation method suggests to choose the following subset

$$S_m^* = \arg\min_{S_m} \max_{\theta\in\Theta(y^N,S_m,p)} \mathrm{DC}(\theta,\hat{\theta}_{S_m}(y^N)) \tag{21}$$

which provides the MDC with validation probability p.

3.2 *Impulse Response Identification of an LTI system*

Finite length input and corrupted output of an LTI system, which is at rest, is available. The system output is corrupted by an additive white Gaussian noise (AWGN) which is zero mean and has variance σ_w^2. The goal is to find the best estimate of the impulse response of the system.

Note that for a system which is at rest the input and output of length N are related to each other only by h^*, the first N elements of the impulse response. Therefore the unknown h^* is an element of a set of order N, S_N ($M=N$). By implementing the MDC we want to choose an estimate of h^* of proper length $m^* \leq N$ which minimizes the description complexity of the true system.

Subset S_m of R^N represents one of the spaces of impulse responses of length m. The input-output relationship of the system is

$$\begin{aligned} y &= \bar{y} + w = h^* * u + w \\ &= h^*_{S_m} * u + \Delta_{S_m} * u + w \\ &= A_{S_m} h^*_{S_m} + B_{S_m} \Delta_m + w \end{aligned} \tag{22}$$

where u is the input, $\bar{y}$ is the noiseless output and y is the noisy output. Also A_{S_m} and B_{S_m} are functions of input u. $h^*_{S_m}$ is the projection of h^* in S_m. It is an element of S_m which is a vector of length N with only m nonzero elements. In each subset S_m, $\hat{h}_{S_m}(y^N)$ is the ML estimate of h.

Here we use the cost function of form (10) where the logarithm is a natural logarithm. The description complexity of random variable Y^N in (5), when the data is generated by h_1,

$$\mathrm{DC}_N(h_1,h_2) = \log\sqrt{2\pi\sigma_w^2} + E_{h_1}\left(\frac{||Y^N - \bar{y}_{h_2}||^2}{2N\sigma_w^2}\right) \tag{23}$$

where $\bar{y}_{h_2} = u * h_2$. Note that in this scenario we have

$$\mathrm{DC}_N(h,h) = H_\theta(Y^N)\log 2 = \log\sqrt{2\pi\sigma_w^2} + \frac{1}{2} \tag{24}$$

which is the same for all elements of S_M. Therefore, the comparison of the DC and comparison of Kullback-Leibler distance in (13) are the same.

In each subset $\bar{h}_{S_m}(N)$, defined in (6), is

$$\begin{aligned}\bar{h}_{S_m}(N) = h_{S_m} + \\ \left(\frac{1}{N}A_{S_m}^T(N)A_{S_m}(N)\right)^{-1}\frac{1}{N}A_{S_m}^T(N)B_{S_m}(N)\Delta_{S_m}.\end{aligned} \tag{25}$$

where h_{S_m} is the projection of h in subset S_m. The minimum description complexity of h in S_m is

$$\begin{aligned}\mathrm{DC}_N(h,\bar{h}_{S_m}(N)) = \log\sqrt{2\pi\sigma_w^2} + \\ \frac{1}{2}(1 + \frac{1}{N\sigma_w^2}||G_{S_m}(N)B_{S_m}(N)\Delta_{S_m}||^2)\end{aligned} \tag{26}$$

where

$$\begin{aligned}G_{S_m}(N) = I - \\ \frac{1}{N}A_{S_m}(N)\left(\frac{1}{N}A_{S_m}^T(N)A_{S_m}(N)\right)^{-1}A_{S_m}^T(N)\end{aligned} \tag{27}$$

is a projection matrix.

3.3 *MDC and order estimation*

The ML estimate in this example is an efficient estimator which satisfies the necessary condition in (14)

$$\hat{h}_{S_m}(y^N) = \arg\min_{g\in S_m}||y^N - y_g||^2 \tag{28}$$

where $y_g = u * g$. The observed data is generated by h^*, the unknown elements of S_M.

The goal is to find probabilistic bounds on $\mathrm{DC}_N(h^*,\hat{h}_{S_m}(y^N))$ based on the observed data in each subset S_m. The first step is the validation step in which $\mathrm{DC}_N(h^*,\bar{h}_{S_m}(N))$ is validated. This calculation is based on the observed error

$$V(\hat{h}_{S_m}(y^N),y^N) = \log\sqrt{2\pi\sigma_w^2} + \frac{1}{2}\left(1 + \frac{||y - \hat{y}_{S_m}||^2}{N\sigma_w^2}\right) \tag{29}$$

where $\hat{y}_{S_m} = u * \hat{h}_{S_m}(y^N)$. This is a sample of a Chi-square random variable with the following expected value and variance

$$E_h(V(\hat{h}_{S_m}(Y^N),Y^N)) = \mathrm{DC}_N\left(h,\bar{h}_{S_m}(N)\right) + \frac{1}{2}\frac{M-m}{N} \tag{30}$$

$$\begin{aligned}\mathrm{var}\left(V(\hat{h}_{S_m}(Y^N),Y^N)\right) = \frac{1}{2N}\frac{M-m}{N} \\ +\frac{1}{N\sigma_w^2}\left(\mathrm{DC}_N\left(h,\bar{h}_{S_m}(N)\right) - \mathrm{DC}_N(h,h)\right)\end{aligned} \tag{31}$$

Therefore, for a chosen p_1, the set $\mathrm{DC}_N(h^*,\bar{h}_{S_m}(N))$ is validated by using the Chi-square distribution table. This set is then used to find bounds on the DC criterion $\mathrm{DC}_N(h^*,\hat{h}_{S_m}(y^N))$ in (21) for each subset. MDC chooses the subset which minimizes the obtained upper bound on the description complexity.

3.4 *Estimation of MDC*

Here we use the properties of the second order statistics of the random variable $V(\hat{h}_{S_m}(Y^N),Y^N))$. The expected value and variance of this random variable is such that the validation step in calculation of $\Theta(y^N,S_m,p)$ can provide bounds on $\mathrm{DC}_N\left(h^*,\bar{h}_{S_m}^*(N)\right)$

$$L_{S_m}(y^N,p_1) \le \mathrm{DC}_N\left(h^*,\bar{h}_{S_m}^*(N)\right) \le U_{S_m}(y^N,p_1) \tag{32}$$

On the other hand $\mathrm{DC}_N(h,\hat{h}_{S_m}(y^N))$ itself is a random variable

$$\begin{aligned}\mathrm{DC}_N(h,\hat{h}_{S_m}(y^N)) = \log\sqrt{2\pi\sigma_w^2} + \\ E_h\left(\frac{||Y^N - \hat{y}_{S_m}||^2}{2N\sigma_w^2}\right)\end{aligned} \tag{33}$$

which is a Chi-square random variable with the following expected value and variance

$$E_h\mathrm{DC}_N(h,\hat{h}_{S_m}(Y^N)) = \mathrm{DC}_N\left(h,\bar{h}_{S_m}(N)\right) + \frac{m}{2N} \tag{34}$$

$$\begin{aligned}\mathrm{var}_h\mathrm{DC}_N(h,\hat{h}_{S_m}(Y^N)) = \frac{m}{2N^2} \\ +\frac{1}{N\sigma_w^2}\left(\mathrm{DC}_N(h,\bar{h}_{S_m}(N)) - \mathrm{DC}_N(h,h)\right)\end{aligned} \tag{35}$$

The second order statistics of this random variable depends only on $\mathrm{DC}_N(h,\bar{h}_{S_m}(N))$, m, N, and $\mathrm{DC}_N(h,h)$, which is fixed for all h. Therefore, by using $\mathrm{DC}_N(h,\bar{h}_{S_m}(N))$ we can provide probabilistic bounds on $\mathrm{DC}_N(h,\hat{h}_{S_m}(y^N))$

$$\begin{aligned}|\mathrm{DC}_N(h,\hat{h}_{S_m}(y^N)) - E_h\mathrm{DC}_N(h,\hat{h}_{S_m}(Y^N))| \le \\ \varepsilon_p(h,S_m,N)\end{aligned} \tag{36}$$

The probability p is the probability that this DC is at most in $\varepsilon_p(h,S_m,N)$ distance of its expected value.

Hence, with probability p_1 bounds on $\mathrm{DC}_N(h^*, \bar{h}^*_{S_m}(N))$ can be *validated* and without calculation of the set $\Theta(y^N, S_m, p)$ *probabilistic* bounds, with probability p, on $\mathrm{DC}_N(h^*, \hat{h}_{S_m}(Y^N))$ can be calculated. The provided bounds are

$$d_L(y^N, S_m, p, p_1) \leq \mathrm{DC}_N(h^*, \hat{h}_{S_m}(y^N)) \leq d_U(y^N, S_m, p, p_1) \quad (37)$$

The optimum subset, using this estimate of MDC, is

$$S^*_m(y^N) = \arg\min_{S_m} d_U(y^N, S_m, p, p_1). \quad (38)$$

When m and $N - m$ are large enough, the Chi-square distributions of $V(\hat{h}_{S_m}(y^N), y^N)$ and $\mathrm{DC}_N\left(h^*, \hat{h}_{S_m}(y^N)\right)$ can be well estimated with Gaussian distributions. In this case the validation probability p_1 and the confidence probability p can be defined in term of $Q(\cdot)$ function [1]. The following theorem provides bounds on the desired DC, $\mathrm{DC}_N\left(h^*, \hat{h}_{S_m}(y^N)\right)$, for this scenario. The calculation of the bounds is similar to the quality evaluation of LTI system estimates in (Beheshti and Dahleh, 2000)

Theorem When m, the order of S_m, and $M - m$ are large enough the Chi-square distributions in of $V(\hat{h}_{S_m}(y^N), y^N)$ and $\mathrm{DC}_N\left(h^*, \hat{h}_{S_m}(y^N)\right)$ can be estimated with Gaussian distributions. Consider $p_1 = Q(\alpha)$ and $p = Q(\beta)$. Then for the LTI system in (22), the upper and lower bounds $d_L(y^N, S_m, p, p_1)$ and $d_U(y^N, S_m, p, p_1)$ are

$$d_U(y^N, S_m, Q(\alpha), Q(\beta)) = \frac{m}{2N} + \frac{1}{2} + \log\sqrt{2\pi\sigma_w^2} + 2\sigma_w^2 U_{S_m} + \frac{\beta}{\sqrt{N}}\sqrt{\frac{m}{2N} + 2U_{S_m}} \quad (39)$$

and

$$d_L(y^N, S_m, Q(\alpha), Q(\beta)) = \max\{0, \frac{m}{2N} + \frac{1}{2} + \log\sqrt{2\pi\sigma_w^2} + 2\sigma_w^2 L_{S_m} - \frac{\beta}{\sqrt{N}}\sqrt{\frac{m}{2N} + 2U_{S_m}}\} \quad (40)$$

where L_{S_m} and U_{S_m} are defined as follows

$$U_{S_m} = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} + K_{S_m}(\alpha). \quad (41)$$

where $m_w = (1 - \frac{m}{N})\sigma_w^2$, and

$$x_{S_m} = \frac{1}{N}||y^N - \hat{y}^N_{S_m}||^2 \quad (42)$$

and

$$K_{S_m}(\alpha) = 2\alpha\frac{\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w}. \quad (43)$$

If $(m_w - \alpha\sqrt{v_m}) \leq x_{S_m} \leq (m_w + \alpha\sqrt{v_m})$,where $v_m = \frac{2}{N}(1 - \frac{m}{N})\sigma_w^4$, the lower bound L_{S_m} is zero and if $(m_w + \alpha\sqrt{v_m}) \leq x_{S_m}$ then

$$L_{S_m} = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} - K_{S_m}(\alpha) \quad (44)$$

Consider the following conditions on α and β

$$\alpha_N \geq \sqrt{\frac{N}{2}}\,(\,1 - \frac{x_m}{(1 - \frac{m}{N})\sigma_w^2}), \quad (45)$$

$$\lim_{N\to\infty} \alpha_N = \infty\,,\ \lim_{N\to\infty} \beta_N = \infty, \quad (46)$$

$$\lim \frac{\alpha_N}{\sqrt{N}} = 0\,,\ \lim_{N\to\infty} \frac{\beta_N}{\sqrt{N}} = 0. \quad (47)$$

These are the sufficient conditions for the bounds on the DCs to approach each other for when $m << N$. Also the conditions guarantee that the validation and confidence probabilities $p_1 = Q(\alpha)$ and $p = Q(\beta)$ approach one as length of data, N, grows.

3.5 *Comparison of Order Estimation Methods*

AIC in (1) and description length in two-stage MDL (2) for the LTI system in (22) are

$$\mathrm{AIC}_{S_m}(y^N) = -\log(\frac{1}{\sqrt{2\pi}\sigma_w}e^{-\frac{||y-\hat{y}_{S_m}||^2}{2N\sigma_w^2}}) + \frac{m}{N} \quad (48)$$

$$\mathrm{DL}_{S_m}(y^N) = -\log(\frac{1}{\sqrt{2\pi}\sigma_w}e^{-\frac{||y-\hat{y}_{S_m}||^2}{2N\sigma_w^2}}) + m\frac{\log N}{2N} \quad (49)$$

Similar to the MDC criterion, these criteria are functions of the output error (42), the variance of the additive noise, length of the data and order of the subset. However, unlike MDC calculation, to calculate these criteria, it is assumed that the true impulse response is an element of the subset S_m!.

In practical applications one important method of order estimation evaluation is to check if the method is consistent. A consistent method is able to point to the subset with smallest order which includes the true model set as the length of the data grows. It is known that MDL is a consistent order estimation method and AIC is not a consistent method. For MDC, the consistency of the method is guaranteed by the proper choice of α, and β. As the length of the data grows these parameters have to be chosen such that the validation and estimation probabilities approach one. Therefore, an improper choice of $\alpha = \beta = 0$ leads to a criterion which is not consistent. It is important to note that for subset S_m which includes the true model set, the MDC criterion in (39) with $\alpha = \beta = 0$ is the AIC in (48). Also, it should be mentioned that the calculated MDC for LTI systems in this paper is the same as the new MDL criterion for linear models which is introduced in (Beheshti and Dahleh, 2003) and is comparable with the two-stage MDL.

[1] $Q(x) = \frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-t^2/2}dt$

When the signal to noise ratio is considerably large and the true system has an infinite length impulse response, the behavior of a consistent method might not be desirable. In this case a practical method should be able to suggest a threshold on the criterion, otherwise the consistent method chooses the model set with the highest order. For this scenario while MDC is able to provide a thresholding method, MDL thresholding is not possible. More detailed discussion on these practical issues is in (Beheshti, 2002) and (Beheshti and Dahleh, 2002).

4. CONCLUSION

In this paper we presented MDC, a new method of order estimation. We elaborated on the advantages of this consistent method over the available information theoretic solutions. It was shown that AIC is a special case of MDC criterion. In this paper the proposed method calculated the description complexity of noisy data for a family of Gaussian distributions. The approach can be extended for calculation of the description complexity for more general classes of linear models with additive noises and also for when the variance of the additive noise is unknown.

5. REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control* **19**, 716–723.

A.P. Liavas, P.A. Regalia and J.Delmas (1999). Blind channel approximation: effective channel order estimation. *IEEE Trans. on Signal Processing* **47**, 3336–3344.

Barron, Y., J. Rissanen and B. Yu (1998). he minimum description length principle in coding and modeling. *IEEE Trans. on Information Theory* **44**, 2743–2760.

Beheshti, S. (2002). *Minimum Description Complexity, Ph.D. thesis*. MIT.

Beheshti, S. and M.A. Dahleh (2000). On model quality evaluation of stable lti systems. *Proceeding of the 39th IEEE Conference on Decision and Control*.

Beheshti, S. and M.A. Dahleh (2002). On denoising and signal representation. *Proceedings of the 10th Mediterranean Conference on Control and Automation*.

Beheshti, S. and M.A. Dahleh (2003). A new minimum description length. *Proceeding of the IEEE Conference on American Control Conference*.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. on Information Theory* **30**, 629–636.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Wax, M. and T. Kailath (1985). Detection of signals by information theoretic criteria. *IEEE Trans. Acoust., Speech, Signal Processing* **33**, 387–392.