A New Minimum Description Length

Soosan Beheshti, Munther A. Dahleh
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
soosan@mit.edu,dahleh@lids.mit.edu

Abstract

The minimum description length(MDL) method is one of the pioneer methods of parametric order estimation with a wide range of applications. We investigate the definition of two-stage MDL for parametric linear model sets and exhibit some drawbacks of the theory behind the existing MDL. We introduce a new description length which is inspired by the Kolmogorov complexity principle.

1 Introduction

"One should not increase, beyond what is necessary, the number of entities required to explain anything"

Occam's razor is a logical principle attributed to the medieval philosopher William of Occam. Applying the principal to the statement above, the main message is

"The simplest explanation is the best"

The principle states that one should not make more assumptions than the minimum needed. A computer scientific approach to this principle is manifested in Kolmogorov complexity. Let y be a finite binary string and let $\mathcal{U}$ be a universal computer. Let $l(y)$ denote the length of the string y. Let $\mathcal{U}(pg)$ denote the output of the computer U when presented with program pg. Then the Kolmogorov complexity $K_{\mathcal{U}}(y)$ of a string y with respect to a universal computer $\mathcal{U}$ is defined as

$$K_{\mathcal{U}}(y) = \min_{pg\,:\,\mathcal{U}(pg)=y} l(pg) \tag{1}$$

The complexity of string y is called the minimum description length of y. For any other computer $\mathcal{A}$ we have

$$K_{\mathcal{U}}(y) \leq K_{\mathcal{A}}(y) + c_{\mathcal{A}} \tag{2}$$

where $c_{\mathcal{A}}$ does not depend on y. This inequality is known as universality of Kolmogorov complexity. Kolmogorov complexity is a modern notion of randomness dealing with the quantity of information in individual objects; that is "pointwise" randomness rather than average randomness produced by a random source.

The two-stage MDL is one of the pioneer methods of computation of description length which is suggested based on this principle [9]. Here we address the drawbacks of the method of calculation of MDL and define a new MDL based on the same Kolmogorov principle. We focus on the order estimation problem for when the data is generated by a linear model with additive noise. The new proposed order estimation method is comparable to other well known order estimation methods such as AIC [1], BIC [10] and other forms of existing MDL methods [2].

2 Problem Statement:Linear Model

Consider the class of parametric model for which the output, $y(n)$, is generated by

$$y(n) = \bar{y}^*(n) + w(n) \tag{3}$$

where $w(n)$ is additive white Gaussian noise with zero mean and variance σ_w^2. Also, $\bar{y}^*$ is the noiseless data. Length N of the data $y^N = [y(1), \cdots y(N)]^T$, a member of random variable set Y^N, is given.

The noiseless data is described with the basis family s_i

$$\bar{y}^*(n) = \sum_{i=1}^{M} s_i(n)\theta^*(i) = A_{S_M}(N)\theta^*_{S_M} \tag{4}$$

where $s_i(n)$ are bounded numbers. Therefore, $A_{S_M}(N)$ is a matrix of dimension $N \times M$. The columns of $A_{S_m}(N)$, s_i's, are independent and we have

$$\frac{1}{N}||s_i||_2^2 \leq c \tag{5}$$

where c is a bounded number and $s_i(n) \neq 0$ for all $1 \leq i \leq M$ and $1 \leq n \leq N$. The value of M can be a part of the prior knowledge of the model class. If it is unknown, the proper assumption of $M = N$ is considered. The parameter $\theta^*_{S_m} = [\theta^*(1), \cdots, \theta^*(M)]$ is a bounded l_1-norm vector in R^M.

A subset S_m of Y^N is of form

$$x^N = [A_{S_m}(N)]\theta_{S_m} \tag{6}$$

where the columns of A_{S_m}, a matrix of dimension $N \times m$, are s_i's which are in S_m $s_i \in S_m$ and θ_{S_m} is a vector of length m for which

$$\theta_{S_m} \in R^m. \tag{7}$$

Inspired by the Kolmogorov complexity and the notion of minimal description length for the string y, we want to search for the subspaces S_m's which can provide the minimum description length(DL) of the "data". Lets first follow principles of the existing two-stage MDL. In each subspace S_m, of order m, the description length of y is described as the minimum codelength which can describe y by an element of S_m. For the codelength in this probabilistic setting the Shannon coding method is used, therefore

$$\mathrm{DL}_{S_m}(y) \quad = \quad \min_{g \in S_m} -\log f(y;g) \tag{8}$$

where the log is based on 2 and $f(Y;g)$ is the probability distribution function(pdf) of random variable Y when the mean is g and the additive noise is w has the same characteristics which were defined for $w[n]$ in (3). Note that in this scenario the probability distribution defined by each g in S_m (and therefore by a θ_{S_m}, $g = A_{S_m}(N)\theta'_{S_m}$) is a Gaussian distribution with output of form

$$x = g + w. \tag{9}$$

The least-square estimate of y in each subspace, which provides the output DL in subspace S_m, is

$$\hat{y}_{S_m} = \arg\min_{g \in S_m} ||g - y||^2. \tag{10}$$

The DL in each subspace is then defined as

$$\mathrm{DL}(y;\hat{y}_{S_m}) = \log\left(\sqrt{2\pi\sigma_w^2}\right)^N + \frac{||\hat{y}_{S_m} - y||^2}{2\sigma_w^2}\log e. \tag{11}$$

which is the description length of the noisy data with an element of S_m, $\hat{y}_{S_m}$. The comparison of this description length for different subspaces always leads to the choice of the S_m with largest possible dimension, S_N, for which the output error is zero. This is not a desired outcome, especially if we know that the unknown true number of parameters which can define $\bar{y}^*$ is less than a given M. To avoid this unwanted outcome, two-stage MDL introduces a codelength which describes elements of S_m as well [9]. Here the assumption is that the length of the code describing any element of subspace S_m is the same and is of order

$$\frac{m}{2}\log(N) \tag{12}$$

Therefore, the total codelength describing y in subspace S_m is defined as the codelength describing elements of S_m in (12) plus the description length of the output given this estimate from (11)

$$\mathrm{DL}(y;S_m) \triangleq \frac{m}{2}\log(N) + \mathrm{DL}(y;\hat{y}_{S_m}). \tag{13}$$

However, choosing the description length for θ_{S_m}s by codes of length $\frac{m}{2}\log(N)$ seems to be an ad-hoc method. Partitioning the subspace of possible θ_{S_m}s can be done with any other descritization per dimension factor other than $\log(N)$. For this reason, it seems that the codelength for all θ_{S_m}s can be of any form $m\log(A\delta)$ when each dimension has $\log A\delta$ elements.

Another method of achieving the description length in (13) is given in [9]. Given a subspace S_m, assume that $M = m$, then

$$\bar{y}^* \in S_m, \tag{14}$$

also assume that the ML estimator in this subspace, $\hat{y}_{S_m}$ approaches the true noiseless data $\bar{y}^*$ as N grows and $\sqrt{N}(\hat{y}_{S_m} - \bar{y}^*)$ converges to a zero mean normal distribution. Then for any prefix code [1] on Y^N, with codelength $L(Y^N)$, the following inequality holds for all prefix codes except a set of prefix codes which will describe later,

$$E\frac{1}{N}L(Y^N) \geq \frac{1}{N}H(f_{\bar{y}^*}(Y^N)) + (1-\epsilon)m\frac{\log(N)}{N}, \tag{15}$$

where $E(X)$ denotes the expected value of random variable X and $H(f_{\bar{y}^*}(Y^N))$ is the differential entropy of Y^N. The inequality holds for all prefix codes except a set of prefix codes which are generated by the pdfs which are generated by a subset of S_m. The Lebesgue measure of this subset goes to zero as N grows [2] [9].

In [9] it is argued that the codelength in (13) is optimum since it can achieve the lower bound in the inequality in (15). However, in [4] it is shown that $m\log(N)$ in the inequality in (15) can be replaced by a family of functions of N, $\log\beta(N)$ for which

$$\lim_{N\to\infty}\beta(N) \quad = \quad \infty \tag{16}$$

$$\lim_{N\to\infty}\frac{1}{(1-\epsilon/2)}\frac{(\beta(N))^{\frac{1-\epsilon}{1-\epsilon/2}}}{N^{m/2}} \quad = \quad 0 \tag{17}$$

Therefore, with the same approach in [9] the codelength in (13) can be generalized to the form

$$\mathrm{DL}(y;S_m) \triangleq \frac{m}{2}\beta(N) + \mathrm{DL}(y;\hat{y}_{S_m}). \tag{18}$$

[1] A code maps the quantized version of elements of Y^N to a set of binary codewords with length $L(y_d^N)$. If no codeword is the prefix of any other, then it is uniquely decidable and is called a *prefix code*

[2] Associated with any pdf on Y^N which is generated by $\bar{y}$ and is nonzero over Y^N, a prefix codelength is available which is proportional to $-\log f(y;\bar{y})$

Hence, not onlylog(N) but any $\beta(N)$ can be used to describe the codelength.

Another important issue is that in practical problems M is an upper bound for m^*, the true number of parameters which generate $\bar{y}^*$, and the challenge is to find the true m^* using the observed data. However, this codelength is provided with the assumption that $\bar{y}^*$ is an element of a given S_m. In practice the same codelength is used for all the subspaces. Even if $\bar{y}^*$ is not an element of S_m. For this subset is this a valid codelength? the answer to this question is not known. The only important fact known about this description length is that it is a consistent criterion. So as the length of data grows if m^* is smaller than M, the method points to the correct m^*. We will discuss this property of MDL more in the following sections.

2.1 New Description Length

The comparison of the codelengths in (11) fails because of the following argument: Minimizing the description length in (8) is the same as

$$\arg \min_{g \in S_m} -\log f(y;g) = \arg \max_{g \in S_m} f(y;g) \tag{19}$$

which provides the ML estimate of $\bar{y}^*$ in each subspace. As we discussed previously, the ML estimation always points to a member of S_N, which has the highest possible order, as a perfect candidate. Therefore, comparison of the codelengths describing y itself in each subspace is not a proper tool for comparison of the estimates $\hat{y}_{S_m}$. Here y is not the string of "data", but y is the data which is corrupted by an additive noise. Therefore the codelength of the ML estimates has to be compared with the pdf which is generated by the "noiseless output", $\bar{y}^*$. To follow the Kolmogorov complexity is to compare the estimate of codelength based on the true, and unknown, pdf. Therefore we consider the following description length in each subspace:

Definition The new description length of "data" in subset S_m is defined as

$$\begin{aligned} \mathrm{DL}(y;S_m) &\triangleq \mathrm{DL}(\hat{y}_{S_m};\bar{y}^*) && (20)\\ &= -\log \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\hat{y}_{S_m}-\bar{y}^*||^2}{2\sigma_w^2}} && (21) \end{aligned}$$

Therefore, the new minimum description length is obtained for S_m^*

$$S_m^* = \arg \min_{S_m} \mathrm{DL}(y;S_m). \tag{22}$$

Calculation of this DL is provided in the following section.

2.2 Calculation of the New Description Length

Calculation and comparison of the description length defined in (21) for different subspaces leads to comparison of the reconstruction error $||\hat{y}_{S_m} - \bar{y}^*||^2$. The available error in each subspace is $||\hat{y}_{S_m} - y||^2$. With this available data we can validate bounds on the description length of $\bar{y}^*_{S_m}$

$$\mathrm{DL}(\bar{y}^*_{S_m};\bar{y}^*) \tag{23}$$

probabilistically, where

$$\bar{y}^*_{S_m} = E(\hat{y}_{S_m}). \tag{24}$$

Details of the validation step is in [5]. For each subspace S_m, the validated upper and lower bounds are functions of order of S_m, m, Length of the data, N, the noise variance, σ_w^2, the data's power and the validation probability P_1. Next step is to provide probabilistic bounds on the desired DL, $\mathrm{DL}(\hat{y}^*_{S_m};\bar{y}^*)$. Probabilistic bounds are also provided in [5]. The bounds are functions of the confidence probability $P2$, Length of data N, noise variance and the validated upper and lower bounds on $\mathrm{DL}(\bar{y}^*_{S_m};\bar{y}^*)$ from step one .

For large N, with P_1 and P_2 approaching one, the upper and lower bounds on the desired description length approach each other and provide a tight estimate for subspaces of low order $m << N$ [5]. Note that in [4] the calculation of another method of order estimation method, minimum description complexity (MDC), is provided. the closed form criterion for MDC and the new MDL have the same structure for the considered linear model class for which the data is generated by the structure given in (3).

3 Thresholding

The existing information theoretic methods attempt to "determine" the true parametric model with m^* parameters. In most practical problems, m^* is not finite and we require to detect the optimum estimate for m^*, which represents the "significant part" of the noise less data $\bar{y}^*$. Implementing the MDL method in this situation, provides an estimate for m^* which is very sensitive to the variation in signal to noise ratio(SNR) [3] and to the length of the output [7]. When the length of the true parameter is infinite, the consistent methods, such as MDL and BIC, point to a higher and higher order as N and/or SNR grows. Some related practical problems of these information theoretic methods are addressed in [11] and [6].

When the true m^* is larger than the length of a very long data, we propose implementing the new information theoretic method of order estimation. With this method we can avoid pointing to higher and higher orders by using a threshold for the description length. If a threshold ϵ is used for the minimum acceptable DL,

[3] SNR$= 10\log_{10} \frac{||Y^N||_2^2}{N\sigma_w^2}$

then we choose the smallest m for which the upper-bound on DL is greater or equal to ϵ. An example of this approach is given in the simulation section.

3.1 MDL Thresholding

Can thresholding be used for the two-stage MDL? In order to make the description length in (13) a valid codelength, which corresponds to a prefix code and satisfies the Kraft's inequality, it is suggested to add a normalizing constant $C(N)$ to the suggested description length

$$\mathrm{DL}_{S_m}(y) = m\log(N) + \log\frac{1}{f_{S_m}(\hat{y}_{S_m};y)} + C_1(N). \quad (25)$$

In [9] it is argued that as N grows $C_1(N)/N \to 0$.

However, note that as N grows the factor $m\frac{\log(N)}{N}$ also goes to zero. For any fixed N, $C_1(N)$ might be comparable with $m\frac{\log(N)}{N}$. Calculation of this normalizing factor is not trivial and is not available. Because of the structure of the DL, in general $C_1(N)$ is a function of the noiseless output $\bar{y}$. Since $C_1(N)$ is a fixed number in the comparison of different subspaces and in calculation of the MDL this term is ignored. However, since $C_1(N)$ changes for different order estimation settings, for example with the change of $\bar{y}^*$, implementation of a threshold is meaningless for this criterion.

Here we prove that the use of threshold is meaningful for the new proposed MDL. Assume that for any problem setting the descritization in output space Y is the same. We prove that the new description length is a codelength of a prefix code. A necessary and sufficient condition for a code to be prefix is that it satisfies the Kraft's inequality. The new DL satisfies the Kraft's inequality by adding a normalized factor. The normalizing factor is not a function of y or $\bar{y}$, but a function of the order of subset S_m and descritization factor of Y. This proves the consistency of the codelengths with universality of Kolmogorov complexity in (2).

Theorem The new description length, defined in (21), satisfies the Kraft's inequality, i.e., corresponds to a prefix code, when a normalized factor $C(N)$ is considered

$$\begin{aligned}\mathrm{DL}_{S_m}(y) &\triangleq \mathrm{DL}(\hat{y}_{S_m};\bar{y}^*) + C(N)\\ &= \log\frac{1}{f_{S_m}(\hat{y}_{S_m};\bar{y}^*)} + C(N). \quad (26)\end{aligned}$$

where

$$C(N) = -\frac{1}{\mathrm{Ln}2}\mathrm{Ln}(\delta^m\sqrt{2\pi\sigma_w^2}^{(N-m)}). \quad (27)$$

Note that although $C(N)$ is a function of m, $C(N)/N$ goes to zero much faster that the terms in the estimate of $\log\frac{1}{f_{S_m}(\hat{y}_{S_m};\bar{y}^*)}$ and it can be ignored for large enough N.

Proof The descritized version of y is y^d and descritized version of $\hat{y}_{S_m}$ is $\hat{y}^d_{S_m}$. The index j is used for the descritized elements in Y^N. Therefore, we have

$$\begin{aligned}&\mathrm{DL}(\hat{y}^d_{S_m}(j);\bar{y}^*) \quad (28)\\ &= -\log\frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N}e^{-\frac{||\bar{y}^*-\hat{y}^d_{S_m}(j)||^2}{2\sigma_w^2}}\\ &= -\frac{1}{\mathrm{Ln}(2)}\mathrm{Ln}\frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N}e^{-\frac{||\bar{y}^*-\hat{y}^d_{S_m}(j)||^2}{2\sigma_w^2}}.\end{aligned}$$

To check the Kraft's inequality for each code word of length $\mathrm{DL}(\hat{y}^d_{S_m}(j);\bar{y}^*)$ we have to show that

$$\sum_j D^{-\mathrm{DL}(\hat{y}^d_{S_m}(j);\bar{y}^*)} \le 1 \quad (29)$$

where D is the size of alphabet resulted from descritizing the output space Y. Equivalently we can check for is added such that

$$\sum_j \left(e^{-\mathrm{DL}(\hat{y}^d_{S_m}(j);\bar{y}^*)}\right)^{\mathrm{LnD}} \le 1 \quad (30)$$

we know that

$$\begin{aligned}&\sum_j \left(e^{-\mathrm{DL}(\hat{y}^d_{S_m}(j);\bar{y}^*)}\right)^{\mathrm{LnD}} \quad (31)\\ &\le \left(\sum_i e^{-\mathrm{DL}(\hat{y}^d_{S_m}(j);\bar{y}^*)}\right)^{\mathrm{LnD}}\\ &\le \left(\sum_j \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N}e^{-\frac{||\bar{y}^*-\hat{y}^d_{S_m}(j)||^2}{2\sigma_w^2}}\right)^{\frac{\mathrm{LnD}}{\mathrm{Ln}2}}\end{aligned}$$

Note that

$$\begin{aligned}&\delta^m\sum_j \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^m}e^{-\frac{||\bar{y}^*-\hat{y}^d_{S_m}(j)||^2}{2\sigma_w^2}} \approx \quad (32)\\ &\int \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^m}e^{-\frac{||\bar{y}^*-\hat{y}_{S_m}||^2}{2\sigma_w^2}}dy\end{aligned}$$

where δ is the precision per dimension in the space Y, or equivalently in the space of the additive noise W. On the other hand, the error $\bar{y}^*_{S_m} - \hat{y}_{S_m}$ has a Gaussian distribution and we have

$$\int \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^m}e^{-\frac{||\bar{y}^*_{S_m}-\hat{y}_{S_m}||^2}{2\sigma_w^2}}dy = 1. \quad (33)$$

Therefore the error $\bar{y}^* - \hat{y}_{S_m}$ also has a Gaussian distribution with same variance and a different mean. Hence, we have

$$\int \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^m}e^{-\frac{||\bar{y}^*-\hat{y}_{S_m}||^2}{2\sigma_w^2}}dy = 1. \quad (34)$$

Therefore

$$\int \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\bar{y}-\hat{y}_{S_m}||^2}{2\sigma_w^2}} dy = \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^{N-m}} \quad (35)$$

Therefore

$$\sum_i \left(e^{-\left(\mathrm{DL}(\hat{y}^d_{S_m}(j);\bar{y}^*) - \frac{1}{\mathrm{Ln}2}\mathrm{Ln}(\delta^m \sqrt{2\pi\sigma_w^2}^{(N-m)})\right)} \right)^{\mathrm{LnD}} \leq 1 \quad (36)$$

Hence, the normalizing factor is $\frac{1}{\mathrm{Ln}2}\mathrm{Ln}(\delta^m \sqrt{2\pi\sigma_w^2}^{(N-m)})$. $\diamondsuit$

4 Linear Models and Simulation Results

Lets consider a linear time invariant(LTI) system for which the matrix $A_{S_m}(N)$ in (3) is a Toeplitz matrix of input. The input of the system is a binary sequence of ± 1. The input is a sample of independent identically distributed Bernoulli random process. Note that in this case the basis s_i in (3) are asymptotically orthogonal. The input-output relationship is of form

$$y(i) = \sum_{k=1}^{i} \theta^*(k) x_{i-k+1} + w(i), \quad (37)$$

We use the microwave radio channel, *chan*10.*mat*, which is available at

http://spib.rice.edu/spib/microwave.html.

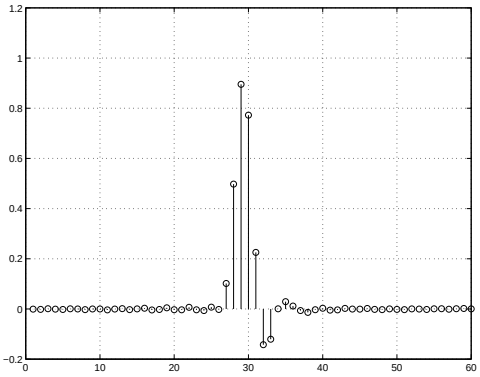

Figure 1: *Real part of the first 60 taps of a microwave radio channel impulse response.*

Figure (1) shows the real part of the first 60 taps of the system impulse response. The simulation result for data of length $N = 300$ and SNR=10db is shown in figure(2). Here the optimum impulse response length for different methods are $\hat{m}$(AIC)=34 and $\hat{m}$(MDL)=32. The new proposed criterion selects $\hat{m} = 33$. Figure(3) shows the upper and lower bound

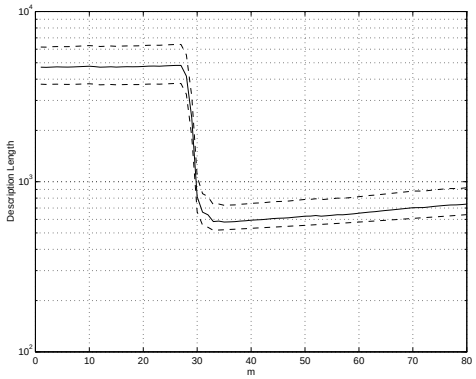

Figure 2: Solid line is the description length for SNR=10db, and N=300. '-.': Probabilistic upperbound and lowerbound of the new description length.

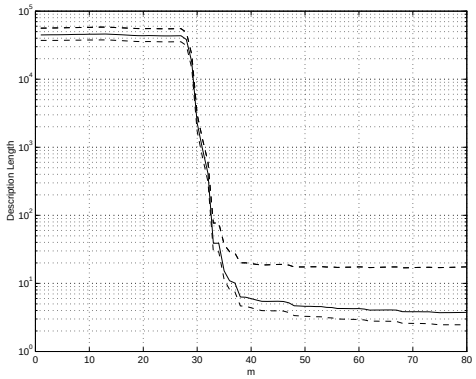

Figure 3: Solid line is the description length for SNR=90db, and N=800. '-.': probabilistic upperbound and lowerbound of the new description length.

on the new DL for N=800, SNR=90db. The bounds on the DL are valid with probability 0.84 and validation probability of 0.84.

In this case all the methods select an impulse response length which is larger than 130. With higher SNR and/or longer data sample, all the methods choose a larger and larger length for the impulse response estimate. However, if we choose a threshold for the DL to be 10, the new criterion selects $m^* = 37$. With this threshold $m^* \leq 37$ when SNR grows and/or the length of data gets larger. Counting for the delay of the system, with the same threshold, the proposed method chooses the 10 taps of the impulse response estimate from 27 to 36 for optimum modelling of the system. .

5 Conclusion

In this paper a new information theoretic method of parametric estimation is introduced. By using the

available data, we are able to probabilistically estimate tight bounds on the new criterion which is in form of a data description length. It is shown that the new description length corresponds to a prefix code and is consistent with the universality of Kolmogorov complexity.

References

[1] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Trans. on Automatic Control,* vol. AC-19, pp.716-723, 1974.

[2] Y. Barron, J. Rissanen, Bin Yu. The Minimum Description Length Principle in Coding and Modeling. *IEEE Trans. on Information theory,* vol 44, pp.2743-2760, Oct. 1998.

[3] S. Beheshti and M.A. Dahleh. On model quality evaluation of stable LTI systems. *Proceedings of the 39th IEEE Conference on Decision and Control,* pp.2716-2721, 2000.

[4] S. Beheshti. *Minimum Description Complexity* . Thesis, MIT, September 2002.

[5] S. Beheshti. and M.A. Dahleh. New Information Theoretic Approach to Order Estimation Problem. *13th IFAC Symposium on System Identification*, August 2003.

[6] W. Chen, K.M.Wong, and J. Reilly. Detection of the Number of Signals: A Predicted Eigen-Threshold Approach. *IEEE Trans on Signal processing,* vol.39, pp.1089-1098, May 1991

[7] A.P. Liavas, P.A. Regalia, and J. Delmas. Blind Channel Approximation: Effective Channel Order Estimation. *IEEE Trans. on Signal Processing,* vol.47, pp.3336-3344, 1999.

[8] L. Ljung. *System Identification: Theory for the user.* NJ: Prentice-Hall, 1998.

[9] J. Rissanen. Universal Coding, Information, Prediction, and Estimation. *IEEE Trans. on Information Theory,* vol.IT30, pp. 629-636, 1984.

[10] G. Schwarz. Estimating The Dimension of a Model. *The Annals of Statistics,* vol.6, pp.461-464, 1978.

[11] K.M.Wong, Q.T. Zhang, J. Reilly and P.C. Yip. On Information Theoretic Criteria for Determining the Number of Signals in High Resolution Array Processing. *IEEE Trans. Acoust. Speech, Signal processing,* vol.38, pp.1959-1971, November 1990