IDIAP RESEARCH REPORT

INTERACTIVE MULTIMODAL INFORMATION MANAGEMENT

EXTENDED BIC CRITERION FOR MODEL SELECTION

Itshak Lapidot Andrew Morris

IDIAP-RR-02-42

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 02-42

EXTENDED BIC CRITERION FOR MODEL SELECTION

Itshak Lapidot Andrew Morris

OCTOBER 2002

Abstract. Model selection is commonly based on some variation of the BIC or minimum message length criteria, such as MML and MDL. In either case the criterion is split into two terms: one for the model (data code length/model complexity) and one for the data given the model (message length/data likelihood). For problems such as change detection, unsupervised segmentation or data clustering it is common practice for the model term to comprise only a sum of sub-model terms. In this paper it is shown that the full model complexity must also take into account the number of sub models and the labels which assign data to each sub model. From this analysis we derive an extended BIC approach (EBIC) for this class of problem. Results with artificial data are given to illustrate the properties of this procedure.

1. Introduction

In model selection by minimum two-part message length encoding, a penalty term is added to the data message length term to account for model complexity. Change detection, segmentation and clustering are unsupervised applications which can apply the BIC, MDL or MML criteria for model selection [1, 2, 3]. In change detection or segmentation it is required to identify change points in a data sequence at which data should be separated and assigned to different models. In clustering unsequenced data must similarly be assigned to some unspecified number of one or more different models. With the BIC model it is assumed that only the number of model parameters needs to be minimized, not the model code length. This model has been found to be successful both in segmentation [2], in clustering [2, 3]. This method, however, usually requires some empirical adjustments, and does not usually take into account the number of clusters, but only the number of parameters in the model for each data cluster. In clustering under a minimum duration constraint, by Ajmera et al [4], the number of model parameters was constant although the number of clusters varied between one and 30, i.e., no penalty was used according to the standard BIC. All these criteria were develop to estimate a model $\mathcal{Q}$ out of a known parametric model class $\mathcal{M}$. In application like clustering and change detection it is required to estimate more than one model from model class $\mathcal{M}$. In this paper it is shown that extra terms for both the number of clusters and the labels which assign the data to each cluster must be added to the usual model code length for optimal model selection.

The principle of two-part minimum message length model selection is briefly presented in Section 2. The proposed extension to the model code length is explained in Section 3. Section 4 presents details of the proposed extension to the BIC message length approximation. Section 5 presents some experiments with artificial data, followed by a discussion and conclusion in Section 6.

2. Two-Part Message Length

Minimum message length model selection is based on the principal that the model which best fits the unseen distribution underlying a given set of model training data is the simplest model which is able to fit the training data to some given level of accuracy. It is very closely related to Bayesian model selection, which selects the model with the maximum posterior probability for the given training data. Model selection uses either one-part or two-part message length. One-part message length is used when the model $\mathcal{Q}$ (defined by a parameter vector $\mathbf{\Theta_M}$, which belongs to a known model class $\mathcal{M}$) is fixed and known to both coder and decoder. In this case the coder only has to send code for the data given the model, $MessLen = code\_length\left(X\left|\mathbf{\Theta_M}\right.\right)$. Two-part message length is used when the model parameters, $\mathbf{\Theta_M}$, are not known to the decoder, so that the coder must estimate and send code for both the model parameters and the data, $MessLen = code\_length\left(\mathbf{\Theta_M}\right) + code\_length\left(X\left|\mathbf{\Theta_M}\right.\right)$, [5, 6]. Like Bayesian model selection [7], minimum message length model selection arises from information theory as an optimal procedure for model selection. In both cases the model code length (model complexity), as well as data code length (data likelihood), must normally be taken into account. Actually MDL [5] and BIC [7] converged to the same formula if we replace the term of data message length by log-likelihood and the model parameters are continuous values that were quantized with uniform distribution over their range.

3. Extended Model Code Length

In MDL and BIC model selection, model complexity is approximated as a simple function of the number of model parameters, $\left|\mathbf{\Theta_M}\right|$. It is easily shown that in clustering, cluster model $\mathcal{S}_\mathbf{K}$ ($\mathbf{\Theta_{M,K}}$ – parameters vector of $\mathcal{S}_\mathbf{K}$) with a mixture model pdf for each of K clusters, and with a fixed combined number M of mixture components, greater K will always result in a greater likelihood, $p\left(X\left|\mathbf{\Theta_{M,K}}=\left\{\mathbf{\Theta_{M_k}}\in\mathcal{M}\right\}_{k=1}^{K}\right.\right)$, and hence smaller data code length. At the same time, the number of model parameters does not change with K. Hence, if model complexity is measured by $\left|\mathbf{\Theta_{M,K}}\right|$ alone, the minimum code length clustering will always use as many clusters as possible, which is absurd. It follows that BIC model selection is not sufficient for data clustering unless some extension to the model structure code length (prior probability) is taken into account, as some function of K, in such way that, when $\left|\mathbf{\Theta_{M,K}}\right|$ is constant, a larger number of clusters results in a higher model complexity.

One can argue that the full definition of cluster model $\mathcal{S}_\mathbf{K}$ requires that the parameter vector $\mathbf{\Theta_{M,K}}$ must be augmented by adding a parameter K to specify the number of clusters, and a set of data labels, $\mathbf{L_K}=\left\{L_n\right\}_{n=1}^{N}$ (N is the size of the data and assumed to be known). To analyze this extended model we should consider two cases. In the first case the data can be rearranged into blocks in the same arbitrary order as the data clusters, but the order of the data within each block is not significant. This would apply, for example, if we wand to code a number of images divided into themed groups. In this case we only need to send the number of data points in each block, $\mathbf{N_K}=\left\{N_k\right\}_{k=1}^{K-1}$, instead of $\mathbf{L_K}$. As the total number of data points is known, then the size of the last block does not need to be included. If we can assume that the probability distributions for K and $\mathbf{N_K}$ (possibly uniform) are known to both coder and decoder, then we must add the following extra terms to the model code length: $code\_length(K)+\sum_{k=1}^{K-1} code\_length(N_k)$. Both terms $code\_length(K)$ and $code\_length(N_k)$ must be non-redundant prefix codes that satisfied the Kraft inequality $\sum_{s_i\in S} 2^{-message\_length(s_i)}\le 1$ (s_i is an element in a set S that represent either K or N_k). Therefore, if we allow the message length to be a fractional, than this quantity is given in terms of log-probabilities as:

$$-\log\left(P(K)\right)-\sum_{k=1}^{K}\log\left(P(N_k)\right) \qquad (1)$$

In the second case the order of the data is important. This case is out the scope of this report and will not be discussed. We only mansion that the simplest solution might be to send all the labels instead of block length and than instead of the term $-\sum_{k=1}^{K-1}\log\left(P(N_k)\right)$, there should present the term $-\sum_{n=1}^{N}\log\left(P(L_n)\right)$.

4. Extended BIC (EBIC) for multi-cluster applications

Given two clustering models based on the same model class $\mathcal{M}$, with parameters $\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}$ and $\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}$ the BIC criterion for choosing which model has the "right" dimension is given in terms of log-likelihood $l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_i}\right.\right)$ and number of parameters $\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_i}\right|$. The criterion to determine which clustering model is better; using BIC is given as follows:

$$l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right.\right) \gtrless l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right.\right) - \frac{\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right| - \left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right|}{2}\log(N) \qquad (2)$$

The chosen model is the one with the higher value. The second term on the right side is the complexity penalty terms according to the difference between numbers of parameters in each model, and the length of the input data, N. In applying this criterion in clustering applications [2, 3] have been found that it is necessary to retrospectively introduce an arbitrary, empirically found, positive scaling factor, λ, for BIC model complexity term.

We now show how equation (2) should be extended to take into account the changing in cluster model complexity term given in equation (1). Let us assume that we have two estimated models from parametric model class $\mathcal{M}$ of *all* mixture models of a given distribution family, such as *all possible* Gaussian mixture model. The model $\mathcal{S}_{\mathbf{K}_i}$ with parameters vector $\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_i}$ has K_i clusters and $M_i = \sum_{k=1}^{K_i} M_{k,i}$ mixture components ($M_{k,i}$ – number of mixture component in cluster k). First consider the case where $M_i = M$. To understand how equations (2) must change it will be sufficient to find the values of $\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right|$ and $\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right|$. For simplicity may assume that the number of parameters of each mixture component is a fixed at R. For a description of the model according to the standard BIC or MDL is required to provide the following number of parameters:

- $M \cdot R$ parameters for all the mixture components in all the clusters.
- Priors of the mixture components in each cluster, $\left\{\left\{P_{m,k,i}\right\}_{m=1}^{M_{k,i}}\right\}_{k=1}^{K_i}$, M parameters.

This gives $\left|\boldsymbol{\Theta}_{\mathbf{M}_i}\right| = M \cdot R + M$, which is independent of i, and the decision is taken only according to the maximum of the likelihoods of the models.

The nature of the parameters of the number of the clusters, K, and the block length, N_k, that are integer values, and they to be different than the $\boldsymbol{\Theta}_{\mathbf{M}}$ parameters, that assumed to be continuous values, and can be analyzed separately in terms of the probabilities associated with each of these integers. If we write the BIC criterion including terms for K_i and $\mathbf{N}_{\mathbf{M},\mathbf{i}}$, than we will get the following:

$$l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right.\right) \gtrless l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right.\right) - \frac{\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right| - \left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right|}{2}\log(N) + \log\left(\frac{P(K_2)}{P(K_1)}\right) + \log\left(\frac{\prod_{k=1}^{K_2} P\left(N_{k,2}\right)}{\prod_{k=1}^{K_1} P\left(N_{k,1}\right)}\right) \qquad (3)$$

In many cases it is reasonable that both K and $N_{k,i}$ are uniformly distributed over finite range, $P(K) = \frac{1}{K_E - K_B + 1}$. and $P\left(N_{k,i}\right) = \frac{1}{N_{\max} - N_{\min} + 1} = \mathrm{P}_{\mathrm{BL}}$. In this case equation (3) becomes the EBIC criterion:

$$l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right.\right)\gtrless l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right.\right)-\frac{\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right|-\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right|}{2}\log\left(N\right)+\left(K_2-K_1\right)\log\left(\mathrm{P}_{\mathrm{BL}}\right) \tag{4}$$

If each segment can be any length in the interval $\left[1,N\right]$ (case of maximum uncertainty), than $\mathrm{P}_{\mathrm{BL}}=\frac{1}{N}$ and the most simplified version of EBIC will be:

$$l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right.\right)\gtrless l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right.\right)-\left(\frac{\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right|-\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right|}{2}+K_2-K_1\right)\log\left(N\right) \tag{5}$$

If $\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right|=\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right|$, than instead of equation (5), the EBIC in will be:

$$l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_1}\right.\right)\gtrless l\left(X\left|\boldsymbol{\Theta}_{\mathbf{M},\mathbf{K}_2}\right.\right)-\left(K_2-K_1\right)\log\left(N\right) \tag{6}$$

$\left(K_2-K_1\right)\log\left(N\right)$ is the maximal penalty. The actual penalty should therefore be scaled as $\varepsilon\left(K_2-K_1\right)\log\left(N\right)$ for some $\varepsilon\in\left(0,1\right]$.

As in BIC, the model penalty is a logarithmic function of N, while the data log-likelihood is proportional to N. The model penalty term therefore more significant for small N and will have negligible effect when N is large.

5. Experiments

Two experiments were conducted to illustrate the effect of EBIC model selection. In the first experiment a data was generated from two Gaussians with the same standard deviations, $\sigma_i=1$ and with expectations $\left\{\mu_{1,t}=0.08t\right\}_{t=1}^{10}$ and $\mu_{2,t}=-\mu_{1,t}$. Two sets were generated: with $N=32$ and $N=128$ points. Each Gaussian generated half of the data. Tests were made under different constrains on the segment length, i.e., it was assumed that several data pointes successively generated from the same source and should be kept together. Segment length were 1, 2, 4, and 8. It should be mentioned that the higher the segment length the less optimal a clustering solution, in terms of log-likelihood, can be achieved. On the other hand more meaningful clusters may be produced. We compare one cluster with two Gaussian mixture components against two clusters with one mixture component each. According to standard BIC no penalty should be used.

Figure 1 shows the result of ΔBIC (if ΔBIC values are less than zero then one cluster is better otherwise two clusters are better). As can be seen a two-cluster model was always better. The black line is the EBIC penalty value for $\varepsilon=1$. It can be seen that there are no big differences between BIC and EBIC except when the ambiguity is high, i.e. when there is a small amount of data, $N=32$, the Gaussians are close one to each other, $\left|\mu\right|<\frac{1}{2}\sigma$ and there is a large duration constraint, $N_s=8$. This indicates that when two clusters are similar EBIC tends to prefer one more accurate cluster with more mixture components.

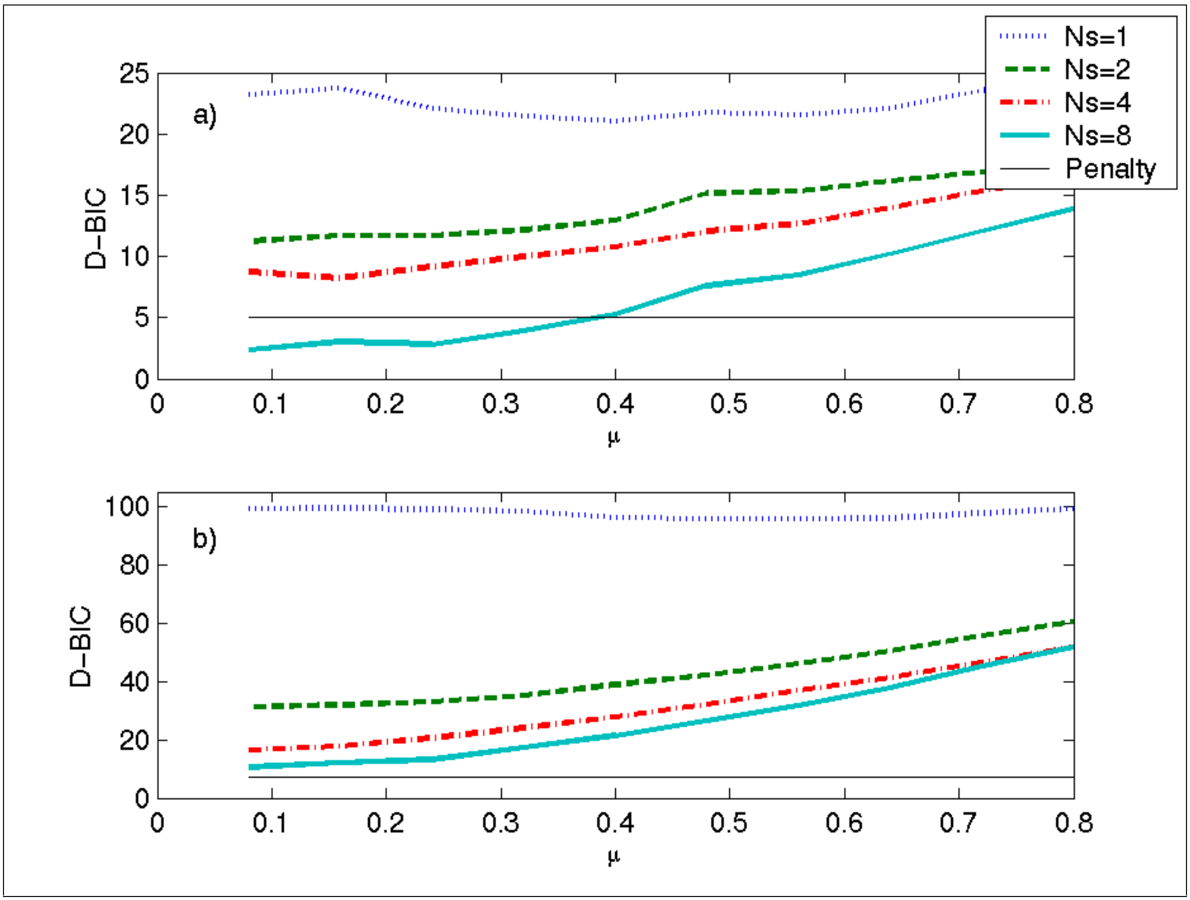

Fig.1: difference between BIC and EBIC for different expectation values, segment length N_s, and amount of data (***a*** - $N = 32$, ***b*** - $N = 128$).

In the second experiment $\mu = 0$ for both Gaussians, $\sigma_1 = 1$, $\sigma_2 = \{0.7, 0.8, 0.9, 1.0\}$ and all the other parameters are as in the first experiment. The results are presented in figure 2. We can see that small μ in the first experiment, and σ_2 close to one in the second experiment, leads the two data sets to have similar statistical properties. So, for small N and large segment length the resulting clusters become similar. While in BIC the decision will be that the two-cluster model is better, EBIC will prefer a one-cluster model.

As was mentioned, a scale factor ε in all the experiments was equal to one. If the scale factor was smaller, the system would be more biased towards the two-cluster model. This parameter can be found empirically (in the same way as the scale factor λ that is used with the BIC criterion), or calculated according to some prior knowledge of another block length distribution.

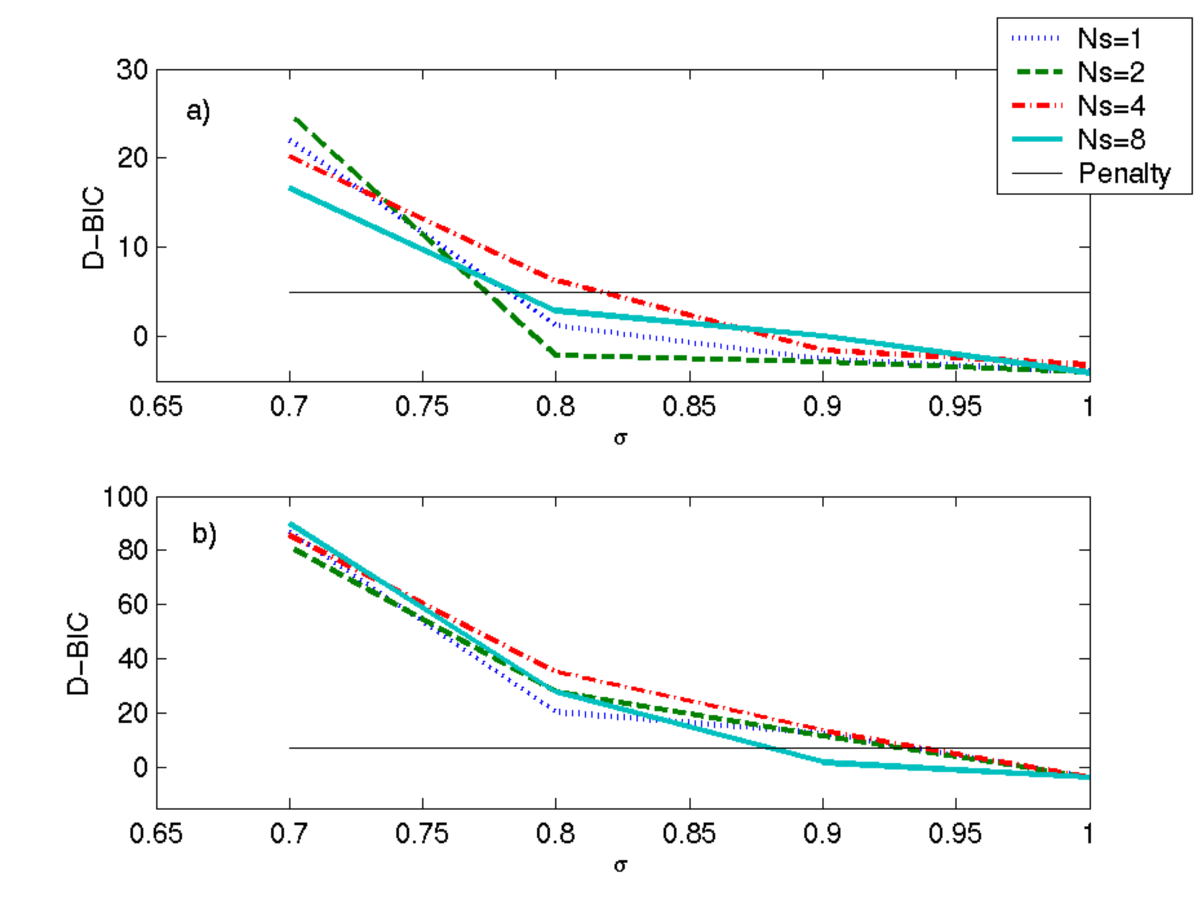

Fig.2: difference between BIC and EBIC for different standard deviation values, segment length N_s, and amount of data (***a*** - $N = 32$, ***b*** - $N = 128$).

6. Discussion

It was shown that the clustering model complexity is not only a function of the number of parameters and their values in the parameter vector $\mathbf{\Theta_{M,K}}$, but also the number of clusters K, and information about the labeling of each data vector $\{L_n\}_{n=1}^{N}$. The labels must not be coded in a direct way, but in a compact way which is just sufficient to permute the data into the blocks as required (in order to minimize the number of parameters to be sent). The code length of such extra information will increase with K.

It was shown that when there is small amount of data or some ambiguity due to the compact nature of the data or clustering constrains, the importance of the additional penalty terms increases.

Acknowledgment

The authors want to thank the Swiss Federal Office for Education and Science (OFES) in the framework of both the EC/OFES MultiModal Meeting Manager (M4) project and the Swiss National Science Foundation through the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2) for supporting this work.

References

[1] J. J. Oliver, R. A. Baxter, and C. S. Wallace, "Unsupervised learning using MML," *Proc. 13th Int. Conf. on Machine Learning*, 1996, pp. 1-10.

[2] M. Cettolo, "Segmentation, classification and clustering of an Italian broadcast news corpus," *Proc. 6th RIAO Conf.*, April 2000, pp. 372-381.

[3] S. S. Chen and P. S. Gapalakrishnan, "Clustering via the Bayesian criterion with applications to speech recognition," *ICASSP'98*, vol. 2, 1998, pp. 645-648.

[4] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM," *ICSLP'02*, USA, 2002, pp.573-576.

[5] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, no. 2, pp. 416-431, 1983.

[6] J. Oliver and R. Baxter, "MML and Bayesianism: similarities and differences: introduction to minimum encoding inference – Part II," Dep. Of Computer Science, Monash University, Clayton, Victoria 3168, Australia, Tech. Rep. TR-206, December 1994.

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