BAYESIAN MODEL COMPARISON AND THE BIC FOR REGRESSION MODELS

Jesper Kjær Nielsen[†], *Mads Græsbøll Christensen*[‡], *Søren Holdt Jensen*[†]

[†]Aalborg University
MISP, Dept. of Electronic Systems
{jkn,shj}@es.aau.dk

[‡]Aalborg University
Audio Analysis Lab, Dept. of Arch., Design & Media Tech.
mgc@create.aau.dk

ABSTRACT

In the signal processing literature, many methods have been proposed for solving the important model comparison and selection problem. However, most of these methods only find the most likely model or only work well under particular circumstances such as a large number of data points or a high signal-to-noise ratio (SNR). One of the most successful classes of methods is the Bayesian information criteria (BIC) and in this paper, we extend some of the recent work on the BIC. In particular, we develop methods in a full Bayesian framework which work well across a large/small number of data points and high/low SNR for either real- or complex-valued data originating from a regression model. Aside from selecting the most probable model, these rules can also be used for model averaging as they assign a probability to each candidate model. Through simulations on a polynomial trend model, we demonstrate that the proposed rules outperform other rules in terms of detecting the true model order, de-noising the noisy signal, and making predictions of unobserved data points. The simulation code is available online.

Index Terms— Model comparison and selection, Bayesian information criterion

1. INTRODUCTION

In many science and engineering applications involving model-based data analysis, the true model structure of the data (if there is one) is often unknown or so complicated that it is intractable to work with. Therefore, a number of simpler models, which are believed to represent the data set accurately, are compared in the light of the data and one, several, or all of these candidate models are used to analyse the data or to make predictions about missing or future data points. Examples of typical model comparison problems in signal processing are to find the number of non-zero regression parameters in linear regression [1, 2], the number of sinusoids in a periodic signal [3, 4], the orders of an autoregressive moving average (ARMA) process [5, 6], and the number of clusters in a mixture model [7, 8]. The model comparison and selection problem is typically formulated in the following way. A data set $\boldsymbol{x} = \begin{bmatrix} x(t_0) & \cdots & x(t_{N-1}) \end{bmatrix}^T$ consisting of either real- or complex-valued numbers is observed, and we assume that these N data points originate from some unknown model. Since we are unsure about the true model, we select a set of K candidate parametric models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K$ which we wish to compare in the light of the data $\boldsymbol{x}$. Here, we assume that the candidate models are regression models of the form

$$\mathcal{M}_k: \quad \boldsymbol{x} = \boldsymbol{s} + \boldsymbol{e} = \boldsymbol{Z}_k \boldsymbol{\alpha}_k + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{s}$ and $\boldsymbol{e}$ form a Wold decomposition of the real- or complex-valued data $\boldsymbol{x}$ into a predictable part and a non-predictable part, respectively. The $N \times l_k$ system matrix $\boldsymbol{Z}_k$ is assumed known whereas the l_k linear parameters in $\boldsymbol{\alpha}_k$ and the noise parameters are unknown. We focus on the regression model for multiple reasons. First, many of the common signal models can be written as or approximated by it [1, 9]. For example, a sinusoidal model can be written as the regression model in (1) after the frequencies have been estimated[1]. Second, the regression model is analytically tractable and therefore results in simple algorithms and facilitates insight into the behaviour of the algorithm.

In the signal processing literature, many methods have been proposed for detecting the most likely model from the set of candidate models [11–13]. This problem is often referred to as model selection and popular examples of model selection rules are the Akaike information criterion (AIC) [14], the Bayesian information criterion (BIC) [15], the asymptotic MAP criteria [9], and many others [16–21]. Contrary to the model selection rules, model comparison methods assign probabilities to all candidate models, and all models (not just the most likely one) can therefore be used to estimate parameters, de-noise the data, and predict future data points. Only a few model comparison methods have been suggested in the signal processing literature (see [12] and the references therein) aside from the Bayesian methods which have been widely studied in the statistical literature [1, 2, 22–24].

1.1. Contributions and Relation to Prior Work

Recently in [25], the authors argue that the BIC is one of the most successful model selection rules when derived properly as recommended in [9]. For the regression model in (1), the BIC is based on analytical approximations of the log-marginal likelihood [25]

$$\ln p(\boldsymbol{x}|\mathcal{M}_k) = -N \ln(\hat{\sigma}^2_{\text{ML}})/r - \ln(|\hat{\boldsymbol{\mathcal{I}}}_k|)/r + \mathcal{O}(1) \tag{2}$$

where $\hat{\sigma}^2_{\text{ML}}$ is the maximum likelihood (ML) estimate of the noise variance, and r is either 1 if $\boldsymbol{x} \in \mathbb{C}^N$ or 2 if $\boldsymbol{x} \in \mathbb{R}^N$. Moreover, $\hat{\boldsymbol{\mathcal{I}}}_k$ is the observed Fisher information matrix (FIM) given by

$$\hat{\boldsymbol{\mathcal{I}}}_k = \begin{bmatrix} \hat{\sigma}^{-2}_{\text{ML}} \boldsymbol{Z}_k^H \boldsymbol{Z}_k & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\sigma}^{-4}_{\text{ML}} N/r \end{bmatrix} . \tag{3}$$

The various forms of the BIC emerge by neglecting first order terms $\mathcal{O}(1)$ and by considering the behaviour of $\hat{\boldsymbol{\mathcal{I}}}_k$ for various values of N and $\hat{\sigma}^2_{\text{ML}}$, and the structure of the system matrix $\boldsymbol{Z}_k$. For example, the most common form of the BIC appears if $\hat{\boldsymbol{\mathcal{I}}}_k$ grows linearly in N so that $\ln(|\hat{\boldsymbol{\mathcal{I}}}_k|) = l_k \ln(N) + \mathcal{O}(1)$. For a polynomial trend model, however, it can be shown that $\ln(|\hat{\boldsymbol{\mathcal{I}}}_k|) = (l_k + 1)^2 \ln(N) + \mathcal{O}(1)$ [9, 25]. In [25], a few new forms of the BIC is derived for the cases of a small/large N and a low/high signal-to-noise ratio (SNR). Based on the work in [2], we here supplement the work in [25] for regression models by developing three methods in a full Bayesian framework

[1] Note that the framework considered in this paper can also be extended to handle non-linear parameters such as the frequency parameters [10].

which we refer to as e-BIC, lp-BIC, and h-BIC. Contrary to the rules suggested in [25], however, the user does not have to decide whether he/she is in a situation with a large/small N and a high/low SNR as this is automatically determined by the e-BIC, the lp-BIC, and the h-BIC. Moreover, there is no need to investigate the behaviour of $\hat{\boldsymbol{\mathcal{I}}}_k$ which might be difficult in some situations such as for non-nested polynomial trend models. This improvement is a consequence of using a full Bayesian framework in which a proper prior distribution is elicited for the linear parameters. Here, we use the g-prior [2] which depends on a single important hyperparameter, and we give a novel physical interpretation of it in terms of the SNR. Another consequence of using a full Bayesian framework is that the e-BIC, the lp-BIC, and the h-BIC can be used for both model selection and comparison. Finally, the computational complexity of the e-BIC and the lp-BIC is similar to that of most other information criteria.

2. MODEL COMPARISON IN REGRESSION MODELS

The e-BIC, the lp-BIC, and the h-BIC are all derived in the same Bayesian framework and are two approximate and one exact solutions to the model comparison problem. Before these rules are derived in Sec 2.2, the Bayesian model is explained which consists of an observation model and prior distributions on the model parameters and models. For model selection and comparison, the elicitation of proper prior distributions on the model parameters is very important as improper prior distributions such as a flat prior on the linear parameters cause the simplest model to be preferred, regardless of the information in the data [10, 22]. This is known as Bartlett's paradox. We therefore give a few arguments for the prior model in Sec. 2.1 and give a new physical interpretation of the important hyperparameter g of the g-prior in Sec. 2.2.1.

2.1. Bayesian Model

2.1.1. The Observation Model

For the non-predictable part $\boldsymbol{e}$, we assume a (complex) normal distribution with probability density function (pdf)

$$p(\boldsymbol{e}|\sigma^2) = \frac{\exp\left(-\frac{\boldsymbol{e}^H\boldsymbol{e}}{r\sigma^2}\right)}{[r\pi\sigma^2]^{N/r}} = \begin{cases}\mathcal{CN}(\boldsymbol{e};\boldsymbol{0},\sigma^2\boldsymbol{I}_N)\ , & r=1\\ \mathcal{N}(\boldsymbol{e};\boldsymbol{0},\sigma^2\boldsymbol{I}_N)\ , & r=2\end{cases} \quad (4)$$

where $(\cdot)^H$ denotes conjugate matrix transposition, and $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. To simplify the notation, we use the non-standard notation $\mathcal{N}_r(\cdot)$ to refer to either the complex normal distribution with pdf $\mathcal{CN}(\cdot)$ for $r = 1$ or the real normal distribution with pdf $\mathcal{N}(\cdot)$ for $r = 2$. Besides being mathematically tractable, arguments such as maximisation of the entropy [26, 27] and the Cramér-Rao bound [28] also favour the white Gaussian noise (WGN) assumption on $\boldsymbol{e}$ [29]. If the noise is known to be coloured, the methods in this paper are still useful if combined with a linear pre-filter. The WGN assumption implies that the observation model is

$$p(\boldsymbol{x}|\boldsymbol{\alpha}_k,\sigma^2,\mathcal{M}_k) = \mathcal{N}_r(\boldsymbol{x};\boldsymbol{Z}_k\boldsymbol{\alpha}_k,\sigma^2\boldsymbol{I}_N)\ . \quad (5)$$

2.1.2. The g-Prior

As the dimension of the vector $\boldsymbol{\alpha}_k$ of linear parameters varies between models, a proper prior distribution must be assigned on it [22]. For regression models, the Zellner's g-prior given by [30]

$$p(\boldsymbol{\alpha}_k|\sigma^2,g,\mathcal{M}_k) = \mathcal{N}_r(\boldsymbol{\alpha}_k;\boldsymbol{0},g\sigma^2[\boldsymbol{Z}_k^H\boldsymbol{Z}_k]^{-1}) \quad (6)$$

has been widely adopted since it leads to analytically tractable marginal likelihoods and is easy to understand and interpret [2]. The g-prior can be interpreted as the posterior distribution on $\boldsymbol{\alpha}_k$ arising from the analysis of a conceptual sample $\boldsymbol{x}_0 = \boldsymbol{0}$ given a uniform prior on $\boldsymbol{\alpha}_k$ and a scaled variance $g\sigma^2$ [31]. The covariance matrix of the g-prior also coincides with a scaled version of the inverse FIM of the linear parameters. Consequently, a large prior variance is therefore assigned to parameters which are difficult to estimate. The noise variance σ^2 is a common parameter and has the same meaning in all models and can therefore be given an improper prior [10, 22]. We therefore use Jeffreys' prior $p(\sigma^2) = (\sigma^2)^{-1}$ which is scale invariant. That is, it includes the same prior knowledge whether we parametrise the model in terms of the noise variance σ^2, the standard deviation σ, or the precision parameter $\lambda = \sigma^{-2}$.

2.1.3. The Models

For the prior on the models, we select a uniform prior of the form $p(\mathcal{M}_k) = K^{-1}\mathbb{I}_{\mathcal{K}}(k)$ where $\mathcal{K} = \{1,\dots,K\}$. However, another prior can easily be used in our framework (see (7) below).

2.2. Bayesian Model Comparison

From Bayes' theorem, we have that the posterior distribution on the models has the probability mass function (pmf)

$$p(\mathcal{M}_k|\boldsymbol{x}) = \frac{\text{BF}[\mathcal{M}_k;\mathcal{M}_b]p(\mathcal{M}_k)}{\sum_{i=1}^{K}\text{BF}[\mathcal{M}_i;\mathcal{M}_b]p(\mathcal{M}_i)} \quad (7)$$

where $\mathcal{M}_b$ is some base model, all other models are compared against, and the Bayes' factor is given by

$$\text{BF}[\mathcal{M}_j;\mathcal{M}_i] = \frac{p(\boldsymbol{x}|\mathcal{M}_j)}{p(\boldsymbol{x}|\mathcal{M}_i)} \triangleq \frac{m_j(\boldsymbol{x})}{m_i(\boldsymbol{x})}\ . \quad (8)$$

The function $m_k(\boldsymbol{x})$ is an unnormalised marginal likelihood whose normalisation constant must be the same for all models. Working with $m_k(\boldsymbol{x})$ rather than the normalised marginal likelihood $p(\boldsymbol{x}|\mathcal{M}_k)$ is usually much simpler. Moreover, $p(\boldsymbol{x}|\mathcal{M}_k)$ does not even exist if an improper prior such as the Jeffreys' prior on the noise variance is used. Given g, the marginal likelihood is given by

$$p(\boldsymbol{x}|g,\mathcal{M}_k) = \int_0^\infty\int_{A_k} p(\boldsymbol{x}|\boldsymbol{\alpha}_k,\sigma^2,\mathcal{M}_k) \\ \times p(\boldsymbol{\alpha}_k|\sigma^2,g,\mathcal{M}_k)p(\sigma^2)d\boldsymbol{\alpha}_k d\sigma^2 \quad (9)$$

where A_k is either the k-dimensional set of real- or complex-valued numbers. By performing the integration in (9), it can be shown that

$$p(\boldsymbol{x}|g,\mathcal{M}_k) \propto m_k(\boldsymbol{x}|g) = \frac{m_N(\boldsymbol{x})}{(1+g)^{l_k/r}}\left(\frac{\hat{\sigma}_N^2}{\hat{\sigma}_k^2(g)}\right)^{N/r} \quad (10)$$

where we have defined

$$\hat{\sigma}_k^2(g) \triangleq \frac{\boldsymbol{x}^H(\boldsymbol{I}_N - \frac{g}{1+g}\boldsymbol{P}_k)\boldsymbol{x}}{N} = \hat{\sigma}_N^2\left(1-\frac{g}{1+g}R_k^2\right) \quad (11)$$

$$R_k^2 \triangleq \frac{\boldsymbol{x}^H\boldsymbol{P}_k\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}} \quad (12)$$

$$m_N(\boldsymbol{x}) \triangleq \Gamma(N/r)(N\pi\hat{\sigma}_N^2)^{-N/r} \quad (13)$$

The matrix $\boldsymbol{P}_k$ is the orthogonal projection matrix of $\boldsymbol{Z}_k$, and $\hat{\sigma}_k^2(g)$ is asymptotically equal to the ML estimate of the noise variance in the limit $\hat{\sigma}_{\text{ML}}^2 = \lim_{g\to\infty}\hat{\sigma}_k^2(g)$. The estimate $\hat{\sigma}_N^2$ is the estimated noise variance of the null model $\mathcal{M}_N$ which is the all-noise model ($l_k = 0$) and has the unnormalised marginal likelihood $m_N(\boldsymbol{x})$. If we select the null model to be the base model, the Bayes' factor given g is thus

$$\mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N|g] = \frac{\left[\hat{\sigma}_N^2/\hat{\sigma}_k^2(g)\right]^{N/r}}{(1+g)^{l_k/r}} = \frac{(1+g)^{(N-l_k)/r}}{(1+g[1-R_k^2])^{N/r}}\ . \tag{14}$$

From (7), we see that the posterior probabilities of the models are proportional to the Bayes' factors for a uniform prior on the models. The most likely model is found by maximising $\mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N|g]$ over k. That is,

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \left[-N\ln\left(\hat{\sigma}_k^2(g)\right) - l_k\ln(1+g)\right] \tag{15}$$

which has the same form as most of the well-known information criteria. From (14) and (15), we also see why the value of g is so vital. In the extreme case of $g\to\infty$ corresponding to a flat prior on the linear parameters, $\mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N|g]\to 0$ for all $\mathcal{M}_k \neq \mathcal{M}_N$, and the null model is therefore always preferred, regardless of the information in the data (Bartlett's paradox). Any finite choice of g will clearly also affect both the estimate of the noise variance and the last term in (15) which is often referred to as the penalty term. However, the value of g is seldomly known in practical applications so we either need to elicit a particular value for it or integrate it out.

2.2.1. *Interpretation of g*

The hyperparameter g can be given a simple physical interpretation in terms of the average SNR and this might be used to select a value for g if one has knowledge of the average SNR in the data. Although a few authors have hinted this connection previously [3,32], the connection established here is to the best of our knowledge new. Define the average SNR of the data as

$$\eta \triangleq \frac{E[\boldsymbol{s}^H\boldsymbol{s}]}{E[\boldsymbol{e}^H\boldsymbol{e}]} = \frac{\boldsymbol{\alpha}_k^H\boldsymbol{Z}_k^H\boldsymbol{Z}_k\boldsymbol{\alpha}_k}{N\sigma^2}\ . \tag{16}$$

Since the random vector $\boldsymbol{y}_k = [(g\sigma^2)^{-1}\boldsymbol{Z}_k^H\boldsymbol{Z}_k]^{1/2}\boldsymbol{\alpha}_k$ is a standard (complex) normal vector when $\boldsymbol{\alpha}_k$ is distributed as in (6), it can be shown that

$$q = \frac{2}{r}\boldsymbol{y}_k^H\boldsymbol{y}_k = \frac{2}{r}\frac{\boldsymbol{\alpha}_k^H\boldsymbol{Z}_k^H\boldsymbol{Z}_k\boldsymbol{\alpha}_k}{g\sigma^2} \tag{17}$$

has a chi-square distribution with pdf $\chi^2(q;2l_k/r)$. Since the SNR is related to q by $\eta = rgq/(2N)$, the prior distribution on the SNR is therefore a gamma distribution with pdf

$$p(\eta|g,\mathcal{M}_k) = \frac{(N/(rg))^{l_k/r}}{\Gamma(l_k/r)}\eta^{l_k/r-1}\exp\left(-\frac{N\eta}{rg}\right) \tag{18}$$

and with mean $E[\eta] = gl_k/N$. Usually, the SNR is expressed in dB by the relation $\tilde{\eta} = 10\log_{10}(\eta)$. Interestingly, $10\log_{10}(E[\eta])$ is the the mode of the pdf on $\tilde{\eta}$. That is, if one knows the SNR in dB of the data, the value of g could be selected as

$$g = (N/l_k)10^{\tilde{\eta}/10}\ . \tag{19}$$

Note that this interpretation also holds for models in which $\boldsymbol{Z}_k$ is parametrised by unknown and non-linear parameters such as sinusoidal frequencies.

2.2.2. *The Emperical BIC (e-BIC)*

If the SNR is unknown, the value of g can also be estimated. A local empirical Bayesian estimate is the maximiser of the marginal likelihood w.r.t. g and given by [2]

$$g_k^{\mathrm{EB}} = \arg\max_{g\in\mathbb{R}^+} m_k(\boldsymbol{x}|g) = \max\left(\frac{NR_k^2 - l_k}{(1-R_k^2)l_k}, 0\right) \tag{20}$$

where $\mathbb{R}^+$ is the set of non-negative real-valued numbers. Inserting (20) in (15) yields the e-BIC for $R_k^2 > l_k/N$ as

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \big[-N\ln\left(\hat{\sigma}_{\mathrm{ML}}^2\right) - l_k\big(\ln(1+g_k^{\mathrm{EB}}) - N\ln(1-l_k/N)/l_k\big)\big] \tag{21}$$

For $N \gg l_k$, the e-BIC is approximately

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \big[-N\ln\left(\hat{\sigma}_{\mathrm{ML}}^2\right) - l_k\big(\ln(1+g_k^{\mathrm{EB}}) + 1\big)\big] \tag{22}$$

From this approximation and the SNR interpretation of g, several interesting observations can be made. When the SNR is large enough to justify that $g_k^{\mathrm{EB}} \gg 1$, the e-BIC is basically a corrected BIC which takes the estimated SNR of the data into account. The penalty coefficient grows with the estimated SNR and the chance of over-fitting thus becomes very low, even under high SNR conditions where the AIC, but also the BIC tend to overestimate the model order [33]. When the estimated SNR on the other hand becomes so low that $g_k^{\mathrm{EB}} \ll 1$, the e-BIC reduces to an AIC-like rule which has a constant penalty coefficient in N. In the extreme case of an estimated SNR of zero, the e-BIC reduces to the so-called no-name rule [11]. Interestingly, empirical studies [9,34] have shown that the AIC performs better than the BIC when the SNR in the data is low, and this is automatically captured by the e-BIC. The e-BIC therefore performs well across all SNR values as we demonstrate in Sec. 3.

2.2.3. *The hyper-BIC (h-BIC) and the Laplace-BIC (lp-BIC)*

Instead of treating g as a fixed quantity, it can also be treated as a random variable and integrated out of the marginal likelihood in (10). For mathematical convenience, we assign the hyper-g prior to g with pdf [2]

$$p(g|\delta) = (\delta/r-1)(1+g)^{-\delta/r}\ ,\quad \delta > r\ . \tag{23}$$

The hyperparameter δ should be selected in the interval $r < \delta \leq 2r$ [2]. Besides having some desirable analytical properties, $p(g|\delta)$ reduces to the Jeffreys' prior and the reference prior when $\delta = r$ [35]. However, since this prior is improper, it can only be used when the prior probability of the null model is zero. Marginalising the marginal likelihood over g yields the Bayes' factor of the h-BIC

$$\begin{aligned}\mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N] &= \int_0^\infty \frac{m_k(\boldsymbol{x}|g)}{m_N(\boldsymbol{x})}p(g|\delta)dg \\ &= \frac{\delta-r}{l_k+\delta-r}\,{}_2F_1\left(\frac{N}{r},1;\frac{l_k+\delta}{r};R_k^2\right)\end{aligned} \tag{24}$$

where ${}_2F_1$ is the Gaussian hypergeometric function [36, p. 314]. When N is large or R_k^2 is very close to one, numerical and computational problems with the evaluation of the Gaussian hypergeometric function may be encountered [37]. From a computational point of view, it may therefore not be advantageous to marginalise (24) w.r.t. g analytically. Instead, the Laplace approximation can be used as a simple alternative. By making the Laplace approximation for the parametrisation $\tau = \ln g$, the Bayes' factor of the lp-BIC can be shown to be [10]

$$\mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N] = \mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N|\hat{g}]\frac{\hat{g}(\delta-r)}{r(1+\hat{g})^{\delta/r}}\sqrt{2\pi\gamma(\hat{g})} \tag{25}$$

where we have defined

$$\beta \triangleq [(N-r)R_k^2 + 2r - k - \delta]/r \tag{26}$$

$$\hat{g} \triangleq \frac{r\beta + \sqrt{r^2\beta^2 + 4r(1-R_k^2)(l_k+\delta-r)}}{(1-R_k^2)(l_k+\delta-r)} \tag{27}$$

$$\gamma(\hat{g}) \triangleq \frac{r}{\hat{g}}\left[\frac{N(1-R_k^2)}{[1+\hat{g}(1-R_k^2)]^2} - \frac{(N-k-\delta)}{(1+\hat{g})^2}\right]^{-1}\ . \tag{28}$$

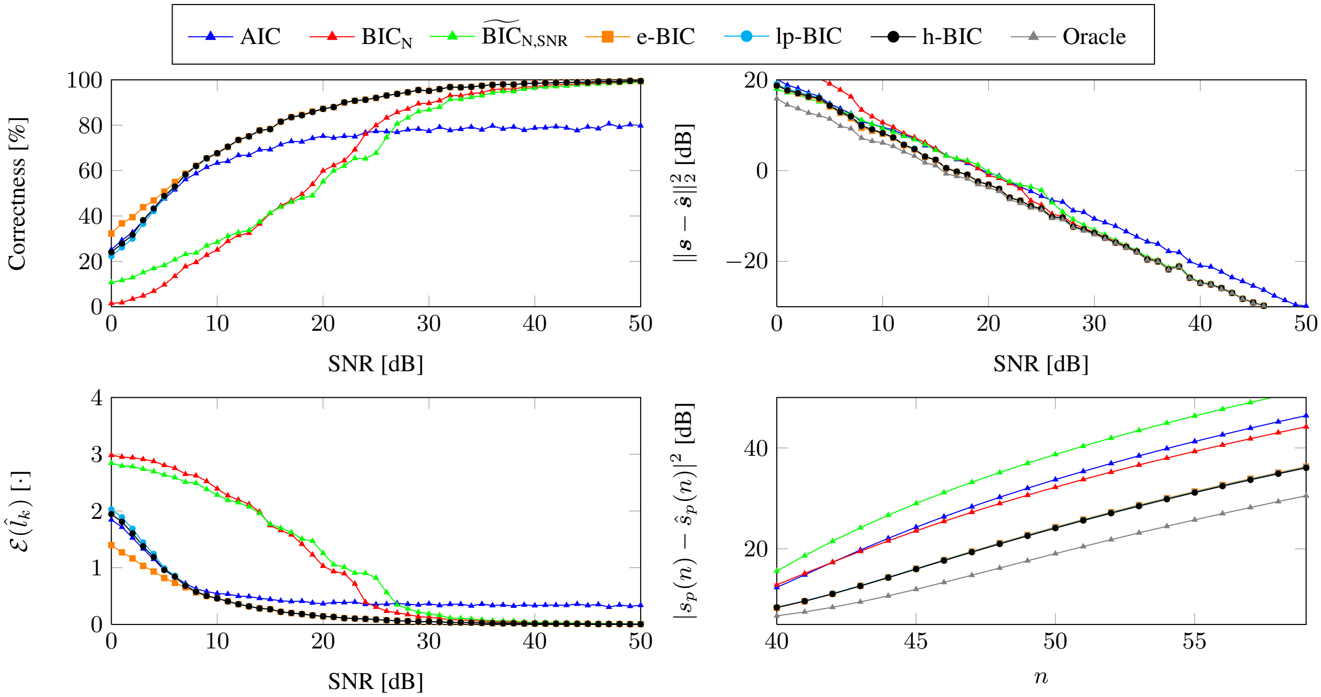

Fig. 1. Performance of various model selection methods. The number of correctly detected models and the MSE of the detected model order are shown in the upper and lower left plots, respectively. In the upper right plot, the denoising performance is shown as a function of the SNR while the lower right plot shows the prediction error as a function of the sample number at an SNR of 15 dB.

Computing the Bayes' factor of the lp-BIC is much faster than computing the Bayes' factor of the h-BIC.

3. SIMULATIONS

Due to the limited space, we cannot present all possible combinations of a large/small N, a high/low SNR, and types of regression models. Therefore, we here present a typical simulation result for a subset of the model selection rules, but encourage the interested reader to try other configurations and rules by modifying the simulation code which is available at `http://kom.aau.dk/~jkn/publications/publications.php`. The simulation presented here is similar to one of the simulations in [25]. We consider a polynomial trend model of a maximum degree of $L = 5$ from which we observe $N = 40$ data points. Although the e-BIC, the lp-BIC, and the h-BIC can handle the situation of $K = 2^L$ models formed by selecting all possible subsets of columns from $\boldsymbol{Z}_k$, we here only consider the $K = L + 1$ nested models since the proper forms of the BIC derived in [25] only holds for this particular case. The performance of the various model selection criteria is evaluated via Monte Carlo simulations consisting of 5000 runs for every SNR which was varied in steps of 1 dB from 0 dB to 50 dB. In contrast to the simulations in [25], however, we do not fix the model and the value for the linear parameters in between runs, but generate a model and the model parameters at random for every run as recommended in [12]. In the simulations, we evaluated the model selection, the denoising, and the prediction performance, and the results are shown in Fig. 1. In the upper left plot, the percentage of correctly detected models is shown. The e-BIC, the lp-BIC, and the h-BIC perform nearly equally well and better than the other rules. The BIC_N and the $\widetilde{\mathrm{BIC}}_{\mathrm{N,SNR}}$ from [25] perform well for high SNRs whereas the AIC performs well at low SNRs. These observations support our claim in Sec 2.2.2. The same observations can be made in the lower left plot where the mean-squared error (MSE) of the detected model order given by $\mathcal{E}(\hat{l}_k) = (l_k - \hat{l}_k)^2$ is shown. In the upper right plot, the de-noising performance is shown. Again we see that the e-BIC, the lp-BIC, and the h-BIC outperform the other rules and have nearly the same de-noising performance as the Oracle who knows the true model order. Finally, the lower right plot shows the prediction performance as a function of the sample number at an SNR of 15 dB. The e-BIC, the lp-BIC, and the h-BIC have nearly identical performance which is only slightly worse than the performance of the Oracle and much better than the performance of the other model selection rules.

4. CONCLUSION

We have here presented three model selection and comparison methods which we have named the e-BIC, the lp-BIC, and the h-BIC. They have all been developed in the same Bayesian framework in which the e-BIC and the lp-BIC can be viewed as approximations to the exact solution h-BIC. The methods differ in how they handle the hyperparameter g of the g-prior which is very important in connection with model selection and comparison. Therefore, we also gave a new physical interpretation of g in terms of the average SNR of the noisy data. Through simulations, the e-BIC, the lp-BIC, and the h-BIC were demonstrated to perform well across a large/small number of data points and a high/low SNR. Moreover, these methods were demonstrated to outperform other methods in terms of model selection, de-noising, and prediction performance.

5. REFERENCES

[1] M. Clyde and E. I. George, “Model uncertainty,” *Statist. Sci.*, vol. 19, no. 1, pp. 81–94, Feb. 2004.

[2] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, “Mixtures of g priors for Bayesian variable selection,” *J. Amer. Statistical Assoc.*, vol. 103, pp. 410–423, Mar. 2008.

[3] C. Andrieu and A. Doucet, “Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC,” *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2667–2676, 1999.

[4] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Morgan & Claypool, 2009.

[5] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, “Reversible jump Markov chain Monte Carlo strategies for Bayesian model selection in autoregressive processes,” *J. of Time Series Analysis*, vol. 25, no. 6, pp. 785–809, Nov. 2004.

[6] T. Cassar, K. P. Camilleri, and S. G. Fabri, “Order estimation of multivariate ARMA models,” *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 3, pp. 494–503, Jun. 2010.

[7] Carl Edward Rasmussen, “The infinite Gaussian mixture model,” in *Adv. in Neural Inf. Process. Syst.*, 2000, pp. 554–560.

[8] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, “Simultaneous feature selection and clustering using mixture models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[9] P. M. Djuric, “Asymptotic MAP criteria for model selection,” *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[10] J. K. Nielsen, *Some New Results on the Estimation of Sinusoids in Noise*, Ph.d. thesis, Aalborg University, Denmark, July 2012.

[11] P. Stoica and Y. Selén, “Model-order selection: a review of information criterion rules,” *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, July 2004.

[12] P. Stoica, Y. Selén, and J. Li, “Multi-model approach to model selection,” *Digital Signal Process.*, vol. 14, no. 5, pp. 399–412, Sep. 2004.

[13] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*, Prentice Hall, May 2005.

[14] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[15] G. Schwarz, “Estimating the dimension of a model,” *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[16] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978.

[17] C. M. Hurvich and C.-L.Tsai, “A corrected Akaike information criterion for vector autoregressive model selection,” *J. of Time Series Analysis*, vol. 14, no. 3, pp. 271–279, May 1993.

[18] J. Rissanen, “A predictive least-squares principle,” *IMA J. Math. Control Inf.*, vol. 3, no. 2–3, pp. 211–222, 1986.

[19] M. G. Christensen, A. Jakobsson, and S. H. Jensen, “Sinusoidal order estimation using angles between subspaces,” *EURASIP J. on Advances in Signal Process.*, vol. 2009, pp. 1–11, Nov. 2009.

[20] R. Badeau, B. David, and G. Richard, “A new perturbation analysis for signal enumeration in rotational invariance techniques,” *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 450–458, Feb. 2006.

[21] J.-M. Papy, L. De Lathauwer, and S. Van Huffel, “A shift invariance-based order-selection technique for exponential data modelling,” *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 473–476, July 2007.

[22] J. O. Berger and L. R. Pericchi, “Objective Bayesian methods for model selection: Introduction and comparison,” *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, vol. 38, pp. 135–207, 2001.

[23] Larry Wasserman, “Bayesian model selection and model averaging,” *J. of Mathematical Psychology*, vol. 44, no. 1, pp. 92–107, Mar. 2000.

[24] A. F. Dentell, *Objective Bayes Criteria for Variable Selection*, Ph.D. thesis, Universitat de Valencia, 2011.

[25] P. Stoica and P. Babu, “On the proper forms of BIC for model order selection,” *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4956–4961, Sept. 2012.

[26] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Springer-Verlag, Berlin Heidelberg, 1988.

[27] E. T. Jaynes, “Bayesian spectrum and chirp analysis,” in *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith and G. J. Erickson, Eds., pp. 1–37. D. Reidel, Dordrecht-Holland, 1987.

[28] P. Stoica and P. Babu, “The Gaussian data assumption leads to the largest Cramér-Rao bound,” *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 132–133, May 2011.

[29] K. Kim and G. Shevlyakov, “Why Gaussianity?,” *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 102–113, Mar. 2008.

[30] A. Zellner, “On assessing prior distributions and Bayesian regression analysis with g-prior distributions,” in *Bayesian Inference and Decision Techniques*. Elsevier, 1986.

[31] D. S. Bové and L. Held, “Hyper-g priors for generalized linear models,” *Bayesian Analysis*, vol. 6, no. 3, pp. 387–410, 2011.

[32] M. Feldkircher and S. Zeugner, “Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in Bayesian model averaging,” Tech. Rep. 09/202, International Monetary Fund, Sept. 2009.

[33] Q. Ding and S. M. Kay, “Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio,” *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1959–1969, May 2011.

[34] Q. T. Zhang and K. M. Wong, “Information theoretic criteria for the determination of the number of signals in spatially correlated noise,” *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1652–1663, Apr. 1993.

[35] R. Guo and P. L. Speckman, “Bayes factor consistency in linear models,” in *The 2009 International Workshop on Objective Bayes Methodology*, Jun. 2009.

[36] I. S. Gradshteĭn, I. M. Ryzhik, and A. Jeffrey, *Table of Integrals, Series, and Products*, Academic Press, 2000.

[37] R. W. Butler and A. T. A. Wood, “Laplace approximations for hypergeometric functions with matrix argument,” *Ann. Stat.*, vol. 30, no. 4, pp. 1155–1177, Aug. 2002.