Bayesian Model Comparison with the g-Prior

Jesper Kjær Nielsen, *Member, IEEE*, Mads Græsbøll Christensen, *Senior Member, IEEE*, Ali Taylan Cemgil, *Member, IEEE*, and Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*—Model comparison and selection is an important problem in many model-based signal processing applications. Often, very simple information criteria such as the Akaike information criterion or the Bayesian information criterion are used despite their shortcomings. Compared to these methods, Djuric's asymptotic MAP rule was an improvement, and in this paper we extend the work by Djuric in several ways. Specifically, we consider the elicitation of proper prior distributions, treat the case of real- and complex-valued data simultaneously in a Bayesian framework similar to that considered by Djuric, and develop new model selection rules for a regression model containing both linear and non-linear parameters. Moreover, we use this framework to give a new interpretation of the popular information criteria and relate their performance to the signal-to-noise ratio of the data. By use of simulations, we also demonstrate that our proposed model comparison and selection rules outperform the traditional information criteria both in terms of detecting the true model and in terms of predicting unobserved data. The simulation code is available online.**

***Index Terms*—Bayesian model comparison, Zellner's g-prior, AIC, BIC, Asymptotic MAP.**

I. Introduction

ESSENTIALLY, all models are wrong, but some are useful [1, p. 424]. This famous quote by Box accurately reflects the problem that scientists and engineers face when they analyse data originating from some physical process. As the exact description of a physical process is usually impossible due to the sheer amount of complexity or an incomplete knowledge, simplified and approximate models are often used instead. In this connection, model comparison and selection methods are vital tools for the elicitation of one or several models which can be used to make inference about physical quantities or to make predictions. Typical model selection problems are to find the number of non-zero regression parameters in linear regression [2]–[4], the number of sinusoids in a periodic signal [5]–[9], the orders of an autoregressive moving average (ARMA) process [10]–[15], and the number of clusters in a mixture model [16]–[18]. For several decades, a large variety of model comparison and selection methods have been developed (see, e.g., [3], [19]–[22] for an overview). These methods can basically be divided in three groups with the first group being those methods which require an a priori estimate of the model parameters, the second group being those methods which do not require such estimates, and the third group being those methods in which the model parameters and model are estimated and detected jointly [15]. The widely used information criteria such as the Akaike information criterion (AIC) [23], the corrected AIC (AIC_c) [24], the generalised information criterion (GIC) [25], the Bayesian information criterion (BIC) [26], the minimum description length (MDL) [27], [28], the Hannan-Quinn information criterion (HQIC) [10], and the predictive least squares [29] belong to the first group of methods. The methods in the second group typically utilise a principal component analysis of the data by analysing the eigenvalues [11], [15], [30], the eigenvectors [31], [32], or the angles between subspaces [33]. In the third group, the Bayesian methods are found. Although these methods are widely used in the statistical community [3], [34]–[37], their use in the signal processing community has only been limited (see, e.g., [7], [8], [14], [38] for a few notable exceptions) compared to the use of the information criteria. The main reasons for this are the high computational costs of running these algorithms and the difficulty of specifying proper prior distributions. A few approximate methods have therefore been developed circumventing most of these issues. Two examples of such approximate methods are the BIC [26] and the asymptotic maximum a posteriori (MAP) rule [39], [40].

The original BIC in [26] and the original MDL principle in [27] are identical in form, but they are derived using very different arguments [22, App. C]. Although this type of rule is one of the most popular model selection methods, it suffers from that every model parameter contributes with the same penalty to the overall model complexity penalty term in the model selection method. Djuric's asymptotic MAP rules [40] improve on this by accounting for that the magnitude of the penalty should depend on the type of models and model parameters being used. For example, the frequency parameter of a sinusoidal signal is shown to contribute with a three times larger penalty term than the sinusoidal amplitude and phase. The asymptotic MAP rules are derived in a Bayesian framework and are therefore sometimes also referred to as Bayesian information criteria [20], [41] when the name alludes to the underlying principle rather than the specific rule suggested in [26][1]. In order to obtain very simple expressions for the asymptotic MAP rules, Djuric uses asymptotic considerations and improper priors, and he also neglects lower order terms during the derivations. The latter is a consequence of the use of improper priors.

J.K. Nielsen and S.H. Jensen are with the Signal and Information Processing Section, Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: {jkn,shj}@es.aau.dk).

M.G. Christensen is with the Audio Analysis Lab, Department of Architecture, Design & Media Technology, Aalborg University, 9220 Aalborg, Denmark (e-mail: mgc@create.aau.dk).

A.T. Cemgil is with the Department of Computer Engineering, Boğaziçi University, 34342 Bebek, Istanbul, Turkey (e-mail: taylan.cemgil@boun.edu.tr).

Digital Object Identifier XXXX

[1]In this paper, the terms MDL and MAP are therefore preferred over BIC.

In this paper, we extend the work by Djuric in several ways. First, we treat the difficult problem of eliciting proper and improper prior distributions on the model parameters. In this connection, we use a prior of the same form as the Zellner's g-prior [42], discuss its properties, and re-parametrise it in terms of the signal-to-noise ratio (SNR) to facilitate a better understanding of it. Second, we treat real- and complex-valued signals simultaneously and propose a few new model selection rules, and third, we derive the most common information criteria in our framework. The latter is useful for assessing the conditions under which the, e.g., AIC and MDL are accurate. As opposed to the various information criteria which are generally derived from cross-validation using the Kullback-Leibler (KL) divergence, we analyse the model comparison problem in a Bayesian framework for numerous reasons [34], [35]; Bayesian model comparison is consistent under very mild conditions, naturally selects the simplest model which explains the data reasonably well (the principle of Occam's razor), takes model uncertainty into account for estimation and prediction, works for non-nested models, enables a more intuitive interpretation of the results, and is conceptually the same, regardless of the number and types of models under consideration. The two major disadvantages of Bayesian model comparison are that the computational cost of running the resulting algorithms may be too high, and that the use of improper and vague prior distributions only leads to sensible answers under certain circumstances. In this paper, we discuss and address both of these issues.

The paper is organised as follows. In Sec. II, we give an introduction to model comparison in a Bayesian framework and discuss some of the difficulties associated with the elicitation of prior distributions and the evaluation of the marginal likelihood. In Sec. III, we propose a general regression model consisting of both linear- and non-linear parameters. For known non-linear parameters, we derive two model comparison algorithms in Sec. IV and give a new interpretation of the traditional information criteria. For unknown non-linear parameters, we also derive a model comparison algorithm in Sec. V. Through simulations, we evaluate the proposed model comparison algorithms in Sec. VI, and Sec. VII concludes this paper.

II. Bayesian Model Comparison

Assume that we observe some real- or complex-valued data

$$\boldsymbol{x} = \begin{bmatrix} x(t_0) & x(t_1) & \cdots & x(t_{N-1}) \end{bmatrix}^T , \tag{1}$$

originating from some unknown model. Since we are unsure about the true model, a set of K candidate parametric models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K$ is elicited to be compared in the light of the data $\boldsymbol{x}$. Each model $\mathcal{M}_k$ is parametrised by the model parameters $\boldsymbol{\theta}_k \in \Theta_k$ where Θ_k is the parameter space of dimension d_k. The relationship between the data $\boldsymbol{x}$ and the model $\mathcal{M}_k$ is given by the probability distribution with density[2] $p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)$ which is called the observation model. When viewed as a function of the model parameters, the observation model is referred to as the likelihood function. The likelihood function plays an important role in statistics where it is used for parameter estimation. However, model selection cannot be solely based on comparing candidate models in terms of their likelihood as a complex model can be made to fit the observed data better than a simple model. The various information criteria are alternative ways of resolving this by introducing a term that penalizes more complex models. This is a manifestation of the well known *Occam's razor* principle which states that if two models explain the data equally well, the simplest model should always be preferred [43, p. 343].

In a Bayesian framework, the model parameters and the model are random variables with the pdf $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$ and pmf $p(\mathcal{M}_k)$, respectively. We refer to these distributions as the prior distributions as they contain our state of knowledge before any data are observed. After observing data, we update our state of knowledge by transforming the prior distributions into the posterior pdf $p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k)$ and pmf $p(\mathcal{M}_k|\boldsymbol{x})$. The prior and posterior distributions for the model parameters and the model are connected by Bayes' theorem

$$p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{p(\boldsymbol{x}|\mathcal{M}_k)} \tag{2}$$

$$p(\mathcal{M}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{M}_k)p(\mathcal{M}_k)}{p(\boldsymbol{x})} \tag{3}$$

where

$$p(\boldsymbol{x}|\mathcal{M}_k) = \int_{\Theta_k} p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k \tag{4}$$

is called the marginal likelihood or the evidence. For model comparison, we often compare the odds of two competing models $\mathcal{M}_j$ and $\mathcal{M}_i$. In this connection, we define the posterior odds which are given by

$$\frac{p(\mathcal{M}_j|\boldsymbol{x})}{p(\mathcal{M}_i|\boldsymbol{x})} = \text{BF}[\mathcal{M}_j; \mathcal{M}_i]\frac{p(\mathcal{M}_j)}{p(\mathcal{M}_i)} \tag{5}$$

where the Bayes' factor is given by

$$\text{BF}[\mathcal{M}_j; \mathcal{M}_i] = \frac{p(\boldsymbol{x}|\mathcal{M}_j)}{p(\boldsymbol{x}|\mathcal{M}_i)} \triangleq \frac{m_j(\boldsymbol{x})}{m_i(\boldsymbol{x})} \tag{6}$$

and $m_k(\boldsymbol{x})$ 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 improper priors are used. We return to this in Sec II-A. Since the prior and posterior distributions of the model are discrete, it is easy to find the posterior odds and the posterior distribution once the Bayes' factors are known. For example, we may rewrite the posterior distribution for the models in terms of the Bayes' factors as

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

where $\mathcal{M}_b$ is some user selected base model which all other models are compared against. Therefore, the main computational challenge in Bayesian model comparison is to compute the unnormalised marginal likelihoods, constituting the Bayes' factor for competing pairs of models. We return to this in Sec II-B. The posterior distribution on the models may be used to select the most probable model. However, as the

[2]In this paper, we have used the generic notation $p(\cdot)$ to denote both a probability density function (pdf) over a continuous parameter and a probability mass function (pmf) over a discrete parameter.

posterior distribution contains the probabilities of all candidate models, all models may be used to make inference about the unknown parameters or to predict unobserved data points. This is called Bayesian model averaging. For example, assume that we are interested in predicting a future data vector $\boldsymbol{x}_p$ using all models. The predictive distribution then has the density

$$p(\boldsymbol{x}_p|\boldsymbol{x}) = \sum_{k=1}^{K} p(\mathcal{M}_k|\boldsymbol{x})p(\boldsymbol{x}_p|\boldsymbol{x},\mathcal{M}_k) \ . \tag{8}$$

Thus, the model averaged prediction is a weighted sum of the predictions from every model.

A. On the Use of Improper Prior Distributions

Like Djuric [39], [40], we might be tempted to use improper prior distributions when we have no or little prior information before observing any data. Whereas this usually works for the inference about model parameters, it usually leads to indeterminate Bayes' factors. To see this, let the prior distribution on the model parameters of the k'th model have the joint density $p(\boldsymbol{\theta}_k|\mathcal{M}_k) = c_k^{-1}h(\boldsymbol{\theta}_k|\mathcal{M}_k)$ where $c_k = \int_{\Theta_k} h(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k$ is the normalisation constant. In the limit $c_k \to \infty$, the prior distribution is said to be improper. An example of a popular improper prior pdf is $h(\boldsymbol{\theta}_k|\mathcal{M}_k) = 1$ so that $p(\boldsymbol{\theta}_k|\mathcal{M}_k) \propto 1$ where $\propto$ denotes proportional to. The posterior distribution on the model parameters has the pdf

$$p(\boldsymbol{\theta}_k|\boldsymbol{x},\mathcal{M}_k) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}_k,\mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{p(\boldsymbol{x}|\mathcal{M}_k)} \tag{9}$$

$$= \frac{p(\boldsymbol{x}|\boldsymbol{\theta}_k,\mathcal{M}_k)h(\boldsymbol{\theta}_k|\mathcal{M}_k)}{\int_{\Theta_k} p(\boldsymbol{x}|\boldsymbol{\theta}_k,\mathcal{M}_k)h(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k} \ . \tag{10}$$

Thus, provided that the integral

$$\tilde{p}(\boldsymbol{x}|\mathcal{M}_k) = \int_{\Theta_k} p(\boldsymbol{x}|\boldsymbol{\theta}_k,\mathcal{M}_k)h(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k \tag{11}$$

converges, the posterior pdf $p(\boldsymbol{\theta}_k|\boldsymbol{x},\mathcal{M}_k)$ is proper even for an improper prior distribution. For two competing models $\mathcal{M}_j$ and $\mathcal{M}_i$, the Bayes' factor is

$$\mathrm{BF}[\mathcal{M}_j;\mathcal{M}_i] = \frac{c_i}{c_j}\frac{\tilde{p}(\boldsymbol{x}|\mathcal{M}_j)}{\tilde{p}(\boldsymbol{x}|\mathcal{M}_i)} \ . \tag{12}$$

The ratio $\tilde{p}(\boldsymbol{x}|\mathcal{M}_j)/\tilde{p}(\boldsymbol{x}|\mathcal{M}_i)$ is well-defined if the posterior distributions on the model parameters $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_i$ are proper. For proper prior distributions, the scalars c_i and c_j are finite, and the Bayes' factor is therefore well-defined. However, for improper prior distributions, the Bayes' factor is in general indeterminate. Specifically, for the improper prior distribution with $h(\boldsymbol{\theta}_j|\mathcal{M}_j) = h(\boldsymbol{\theta}_i|\mathcal{M}_i) = 1$, it can be shown that [44]

$$\frac{c_i}{c_j} = \begin{cases} 0 \ , & d_j > d_i \\ 1 \ , & d_j = d_i \\ \infty \ , & d_j < d_i \end{cases} \tag{13}$$

where d_j and d_i are the number of model parameters in $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_i$, respectively. That is, the simplest model is always preferred over more complex models, regardless of the information in the data. This phenomenon is known as the *Bartlett's paradox*[3] [45]. Due to the Bartlett's paradox, the general rule is that one should use proper prior distributions for model comparison. However, there exists one important exception to this rule which we consider below. From (12), we also see that vague prior distributions may give misleading answers. For example, a vague distribution such as the normal distribution with a very large variance leads to an arbitrary large normalising constant c_k which strongly influences the Bayes' factor [35]. Therefore, the elicitation of proper prior distributions is very important for Bayesian model comparison.

1) Common Model Parameters: Consider the case where one model, the null model $\mathcal{M}_N$, is a sub-model[4] of all other candidate models. That is $\mathcal{M}_N \subseteq \mathcal{M}_k$ for $k = 1,\ldots,K$. We denote the null model parameters as $\boldsymbol{\theta}_N$ and the model parameters of the k'th model as $\boldsymbol{\theta}_k = \begin{bmatrix}\boldsymbol{\theta}_N^T & \boldsymbol{\psi}_k^T\end{bmatrix}^T$ where $(\cdot)^T$ denotes matrix transposition. The prior distribution on $\boldsymbol{\theta}_k$ now has the pdf

$$p(\boldsymbol{\theta}_k|\mathcal{M}_k) = p(\boldsymbol{\psi}_k|\boldsymbol{\theta}_N,\mathcal{M}_k)p(\boldsymbol{\theta}_N|\mathcal{M}_k) \ . \tag{14}$$

If the null model parameters $\boldsymbol{\theta}_N$ and the additional parameters $\boldsymbol{\psi}_k$ are orthogonal[5], then knowledge of the true model does not change the knowledge about $\boldsymbol{\theta}_N$, and we therefore have that $p(\boldsymbol{\theta}_N|\mathcal{M}_k) = p(\boldsymbol{\theta}_N|\mathcal{M}_N)$ [35], [37]. Thus, using the prior pdf $p(\boldsymbol{\theta}_N|\mathcal{M}_N) = c_b^{-1}h(\boldsymbol{\theta}_N|\mathcal{M}_N)$, the Bayes' factor is

$$\mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N] = \frac{\int_{\Theta_k} p(\boldsymbol{x}|\boldsymbol{\theta}_k,\mathcal{M}_k)p(\boldsymbol{\psi}_k|\boldsymbol{\theta}_N,\mathcal{M}_k)h(\boldsymbol{\theta}_N|\mathcal{M}_N)d\boldsymbol{\theta}_k}{\int_{\Theta_b} p(\boldsymbol{x}|\boldsymbol{\theta}_N,\mathcal{M}_N)h(\boldsymbol{\theta}_N|\mathcal{M}_N)d\boldsymbol{\theta}_N} \tag{16}$$

which is proper if the posterior distribution on the null model parameters and the prior distribution with pdf $p(\boldsymbol{\psi}_k|\boldsymbol{\theta}_N,\mathcal{M}_k)$ are proper. That is, the Bayes' factor is well-defined since $c_b = c_i = c_j$ even if an improper prior distribution is selected on the null model parameters, provided that they are orthogonal to the additional model parameters $\boldsymbol{\psi}_k$.

B. Computing the Marginal Likelihood

As alluded to earlier, the main computational difficulty in computing the posterior distribution on the models is the evaluation of the marginal likelihood in (4). The integral may not have a closed-form solution, and direct numerical evaluation may be infeasible if the number of model parameters is too large. Numerous solutions to this problem have been proposed and they can broadly be dichotomised into stochastic methods and deterministic methods. In the stochastic methods, the integral is evaluated using numerical sampling which are also known as Monte Carlo techniques [46]. Popular techniques are importance sampling [47], Chib's methods [48], [49], reversible jump Markov chain Monte Carlo [50], and population Monte Carlo [51]. An overview over and comparison of several methods are given in [52]. An advantage of the stochastic methods is that they in principle

[3]Bartlett's paradox is also called the Lindley's paradox, the Jeffreys' paradox, and various combinations of the three names.

[4]Instead of the null model, the full model, which contains all other candidate models, can also be used [4].

[5]If one set of parameters $\boldsymbol{\theta}_N$ is orthogonal to another set of parameters $\boldsymbol{\psi}_k$, the Fisher information matrix of the joint parameter vector $\boldsymbol{\theta}_k = \begin{bmatrix}\boldsymbol{\theta}_N^T & \boldsymbol{\psi}_k^T\end{bmatrix}^T$ is diagonal. That is,

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_k) = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_N,\boldsymbol{\psi}_k) = \begin{bmatrix}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_N) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mathcal{I}}(\boldsymbol{\psi}_k)\end{bmatrix} \ . \tag{15}$$

can generate exact results. However, it might be difficult to assess the convergence of the underlying stochastic integration algorithm. On the other hand, the deterministic methods can only generate approximate results since they are based on analytical approximations which make the evaluation of the integral in (4) possible. These methods are also sometimes referred to as variational Bayesian methods [53], and a simple and widely used example of these methods is the Laplace approximation [54]. In order to derive the original BIC and the asymptotic MAP rule and since the Laplace approximation is used later in this paper, we briefly describe it here.

1) The Laplace Approximation: Denote the integrand of an integral such as in (4) by $f(\boldsymbol{\xi}_k)$ where $\boldsymbol{\xi}_k = \begin{bmatrix}\mathrm{Re}(\boldsymbol{\theta}_k^T) & \mathrm{Im}(\boldsymbol{\theta}_k^T)\end{bmatrix}^T$ is a vector of $\tilde{d}_k$ real parameters with support Ξ_k. Moreover, suppose there exists a suitable one-to-one transformation $\boldsymbol{\xi}_k = \boldsymbol{h}(\boldsymbol{\varphi}_k)$ such that the logarithm of the integrand

$$q(\boldsymbol{\varphi}_k) = \left|\frac{\partial \boldsymbol{h}(\boldsymbol{\varphi}_k)}{\partial \boldsymbol{\varphi}_k}\right| f\left(\boldsymbol{h}(\boldsymbol{\varphi}_k)\right) \tag{17}$$

can be accurately approximated by the second-order Taylor expansion around a mode $\hat{\boldsymbol{\varphi}}_k$ of $q(\boldsymbol{\varphi}_k)$. That is,

$$\ln q(\boldsymbol{\varphi}_k) \approx \ln q(\hat{\boldsymbol{\varphi}}_k) + \frac{1}{2}(\boldsymbol{\varphi}_k - \hat{\boldsymbol{\varphi}}_k)^T \boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)(\boldsymbol{\varphi}_k - \hat{\boldsymbol{\varphi}}_k) \tag{18}$$

where

$$\boldsymbol{H}(\boldsymbol{\varphi}_k) = \frac{\partial^2 \ln q(\boldsymbol{\varphi}_k)}{\partial \boldsymbol{\varphi}_k \partial \boldsymbol{\varphi}_k^T} \tag{19}$$

is the Hessian matrix. Under certain regularity conditions [40], the Laplace approximation is then given by

$$\int_{\Phi_k} q(\boldsymbol{\varphi}_k) d\boldsymbol{\varphi}_k \approx q(\hat{\boldsymbol{\varphi}}_k)(2\pi)^{\tilde{d}_k/2} |-\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)|^{-1/2} \tag{20}$$

where Φ_k is the support of $\boldsymbol{\varphi}_k$. The main difficulty in computing the Laplace approximation is to find a suitable parametrisation of the integrand so that the second-order Taylor expansion of $\ln q(\boldsymbol{\varphi}_k)$ is accurate. If $q(\boldsymbol{\varphi}_k)$ consists of multiple, significant, and well-separated peaks, an integral can be approximated by a Laplace approximation to each peak at their respective modes [55].

2) The Original BIC and the Asymptotic MAP: The original BIC [26] and the asymptotic MAP rule [40] are based on the Laplace approximation with $\boldsymbol{h}(\cdot)$ being the identity function so that

$$f(\boldsymbol{\xi}_k) = q(\boldsymbol{\varphi}_k) = p(\boldsymbol{x}|\boldsymbol{\xi}_k, \mathcal{M}_k) p(\boldsymbol{\xi}_k|\mathcal{M}_k) \ . \tag{21}$$

By neglecting terms of order $\mathcal{O}(1)$ and assuming a flat prior around $\hat{\boldsymbol{\xi}}_k$, the marginal likelihood in the asymptotic MAP rule is

$$\int_{\Xi_k} f(\boldsymbol{\xi}_k) d\boldsymbol{\xi}_k \approx p(\boldsymbol{x}|\hat{\boldsymbol{\xi}}_k, \mathcal{M}_k) |-\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)|^{-1/2} \ . \tag{22}$$

In the MAP rule, the determinant of the observed information matrix $-\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)$ is evaluated using asymptotic considerations, and the asymptotic result therefore depends on the specific structure of $\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)$, the number of data points, and the SNR [41]. For the original BIC, however, this determinant is assumed to grow linearly in the sample size N so that

$$|-\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)| = \left|-\frac{N}{\alpha}\frac{\alpha}{N}\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)\right| = \left(\frac{N}{\alpha}\right)^{\tilde{d}_k} \mathcal{O}(1) \tag{23}$$

where α is an arbitrary constant. In the original BIC, $\alpha = 1$ and the original BIC is therefore

$$\int_{\Xi_k} f(\boldsymbol{\xi}_k) d\boldsymbol{\xi}_k \approx p(\boldsymbol{x}|\hat{\boldsymbol{\xi}}_k, \mathcal{M}_k) N^{-\tilde{d}_k/2} \ , \tag{24}$$

but α can be selected arbitrarily which we find unsatisfactory. In [40], Djuric shows that the MAP rule and the orignal BIC/MDL coincide for autoregressive models and sinusoidal models with known frequencies. However, he also shows that that they differ for polynomial models, sinusoidal models with unknown frequencies, and chirped signal models.

III. Model Comparison in Regression Models

Bayesian model comparison as outlined in Sec. II is applicable to any model, but we have to work with a specific model to come up with specific algorithms for model comparison. In the rest of this paper, we therefore focus on regression models of the form

$$\mathcal{M}_k : \ \boldsymbol{x} = \boldsymbol{s}_k(\boldsymbol{\phi}_k, \boldsymbol{\psi}, \boldsymbol{\alpha}_k) + \boldsymbol{e} = \boldsymbol{B}\boldsymbol{\psi} + \boldsymbol{Z}_k(\boldsymbol{\phi}_k)\boldsymbol{\alpha}_k + \boldsymbol{e} \tag{25}$$

where $\boldsymbol{s}_k(\boldsymbol{\phi}_k, \boldsymbol{\psi}, \boldsymbol{\alpha}_k)$ 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. Since the model parameters are treated as random variables, the predictable part $\boldsymbol{s}_k(\boldsymbol{\phi}_k, \boldsymbol{\psi}, \boldsymbol{\alpha}_k)$ is also stochastic like the non-predictable part. All models include the same null model

$$\mathcal{M}_N : \ \boldsymbol{x} = \boldsymbol{B}\boldsymbol{\psi} + \boldsymbol{e} \tag{26}$$

where $\boldsymbol{B}$ and $\boldsymbol{\psi}$ are a known $N \times l_N$ system matrix and a known or unknown vector of l_N linear parameters, respectively. Usually, the predictable part of the null model is either taken to be a vector of ones so that $\boldsymbol{\psi}$ acts as an intercept or not present at all. In the latter case, the null model is simply the noise-only model. The various candidate models differ in terms of the l_k linear parameters in the vector $\boldsymbol{\alpha}_k$ and the $N \times l_k$ system matrix $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$, which is parametrised by the ρ_k real-valued and non-linear parameters in the vector $\boldsymbol{\phi}_k$. These non-linear parameters may be either known, unknown or not present at all. We discuss the first and latter case in Sec. IV and the case of unknown non-linear parameters in Sec. V. Without loss of generality, we assume that the columns of $\boldsymbol{B}$ and $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$ are orthogonal to each other so that $\boldsymbol{\psi}$ has the same interpretation in all models and therefore can be assigned an improper prior if $\boldsymbol{\psi}$ is unknown. If the columns of $\boldsymbol{B}$ and $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$ are not orthogonal to each other, $\boldsymbol{s}(\boldsymbol{\phi}_k, \boldsymbol{\psi}, \boldsymbol{\alpha}_k)$ can be re-parametrised so that the columns of the two system matrices are orthogonal [56]. We focus on the regression model in (25) for several reasons. First of all, many common signal models used in signal processing can be written in the form of (25). Examples of such models are the linear regression model, the polynomial regression model, the autoregressive signal model, the sinusoidal model, and the chirped signal model, and these signal models were also considered by Djuric in [40]. Second, the regression model in (25) is analytically tractable and therefore results in computational algorithms with a tractable complexity. Moreover, the analytical tractability facilitates insight into, e.g., the various information criteria. Finally, the regression model in (25) can be viewed as a approximation to more complex models [3].

A. Elicitation of Prior Distributions

In the Bayesian framework, the unknown parameters are random variables. In addition to specifying a distribution on the noise vector, we therefore also have to elicit prior distributions on these unknown parameters. The elicitation of prior distributions is a controversial aspect in Bayesian statistics as it is often argued that subjectivity is introduced into the analysis. We here take a more practical view at this philosophical problem and consider the elicitation as a consistent and explicit way of stating our assumptions. In addition to the philosophical issue, we also face two practical problems in the context of eliciting prior distributions for model comparison. First, if we assume that $l_k \leq L$, we can select a subset of columns from $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$ in $K = 2^L$ different ways. A careful elicitation of the prior distribution for the model parameters in each model is therefore infeasible if L is too large, and we therefore prefer to do the elicitation in a more generic way. Second, even if we have only a vague prior knowledge, the use of improper or vague prior distributions in an attempt to be objective may lead to bad or non-sensible answers [35]. As we discussed in Sec. II, this approach usually works for making inference about model parameters, but may lead to the Bartlett's paradox for model selection.

1) The Noise Distribution: In order to deduce the observation model, we have to select a model for the non-predictable part $\boldsymbol{e}$ of the model in (25). As it is purely stochastic, it must have zero mean, and we assume that it has a finite variance. As advocated by Jaynes and Bretthorst [57]–[59], we select the distribution which maximises the entropy under these constraints. It is well-known, that this distribution is the (complex) normal distribution with pdf

$$p(\boldsymbol{e}|\sigma^2) = \left[r\pi\sigma^2\right]^{-N/r} \exp\left(-\frac{\boldsymbol{e}^H\boldsymbol{e}}{r\sigma^2}\right) \tag{27}$$

$$= \begin{cases} \mathcal{CN}(\boldsymbol{e}; \boldsymbol{0}, \sigma^2\boldsymbol{I}_N) \ , & r = 1 \\ \mathcal{N}(\boldsymbol{e}; \boldsymbol{0}, \sigma^2\boldsymbol{I}_N) \ , & r = 2 \end{cases} \tag{28}$$

where $(\cdot)^H$ denotes conjugate matrix transposition, $\boldsymbol{I}_N$ is the $N \times N$ identity matrix, and r is either 1 if $\boldsymbol{x}$ is complex-valued or 2 if $\boldsymbol{x}$ is real-valued. 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 $r = 2$. It is important to note that the noise variance σ^2 is a random variable. As opposed to the case where it is simply a fixed but unknown quantity, the noise distribution marginalised over this random noise variance is able to model noise with heavy tails and is robust towards outliers. Another important observation is that (28) does not explicitly model any correlations in the noise. However, including correlation constraints into the elicitation of the noise distribution lowers the entropy of the noise distribution which is therefore more informative [58, Ch. 7], [59]. This leads to more accurate estimates when there is genuine prior information about the correlation structure. However, if nothing is known about the correlation structure, the noise distribution in (28) is the best choice since it is the least informative distribution and is thus able to capture every possible correlation structure in the noise [59], [60].

The Gaussian assumption on the noise implies that the observed data are distributed as

$$p(\boldsymbol{x}|\boldsymbol{\alpha}_k, \boldsymbol{\psi}, \boldsymbol{\phi}_k, \sigma^2, \mathcal{M}_k) = \mathcal{N}_r(\boldsymbol{x}; \boldsymbol{B}\boldsymbol{\psi} + \boldsymbol{Z}_k(\boldsymbol{\phi}_k)\boldsymbol{\alpha}_k, \sigma^2\boldsymbol{I}_N) \ . \tag{29}$$

The Fisher information matrix (FIM) for this observation model is derived in App. A and given by (79). The block diagonal structure of the FIM means that the common parameters $\boldsymbol{\psi}$ and σ^2 are orthogonal to the additional model parameters and can therefore be assigned improper prior distributions.

2) The Noise Variance: Since the noise variance is a common parameter in all models and orthogonal to all other parameters, it can be assigned an improper prior. The Jeffreys' prior $p(\sigma^2) = (\sigma^2)^{-1}$ is a widely used improper prior for the noise variance which we also adopt in this paper. The popularity primarily stems from that the prior is invariant under transformations of the form σ^m for all $m \neq 0$. Thus, the Jeffreys' prior includes the same prior knowledge whether we parametrise our model in terms of the noise variance σ^2, the standard deviation σ, or the precision parameter $\lambda = \sigma^{-2}$.

3) The Linear Parameters: Since we have assumed that $\boldsymbol{B}^H\boldsymbol{Z}_k(\boldsymbol{\phi}_k) = \boldsymbol{0}$, the linear parameters $\boldsymbol{\psi}$ of the null model are orthogonal to the remaining parameters. We can therefore use the improper prior distribution with pdf $p(\boldsymbol{\psi}) \propto 1$ for $\boldsymbol{\psi}$. This prior is often used for location parameters as it is translation invariant. As the dimension of the vector $\boldsymbol{\alpha}_k$ of linear parameters varies between models, a proper prior distribution must be assigned on it. For linear regression models, the Zellner's g-prior given by [42]

$$p(\boldsymbol{\alpha}_k|g, \sigma^2, \boldsymbol{\phi}_k, \mathcal{M}_k) = \mathcal{N}_r(\boldsymbol{\alpha}_k; \boldsymbol{0}, g\sigma^2[\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Z}_k(\boldsymbol{\phi}_k)]^{-1}) \tag{30}$$

has been widely adopted since it leads to analytically tractable marginal likelihoods and is easy to understand and interpret [4]. 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 the non-linear parameters $\boldsymbol{\phi}_k$, a uniform prior on $\boldsymbol{\alpha}_k$, and a scaled variance $g\sigma^2$ [61]. Given $\boldsymbol{\phi}_k$, the covariance matrix of the g-prior also coincides with a scaled version of the inverse Fisher information matrix. Consequently, a large prior variance is therefore assigned to parameters which are difficult to estimate. We can also make a physical interpretation of the scalar g when the null model is the noise-only model. In this case, the mean of the prior on the average signal-to-noise ratio (SNR) is [62]

$$E[\eta|g, \mathcal{M}_k] = l_k g/N \ . \tag{31}$$

Moreover, this value is also the mode of the prior on the average SNR in dB [62].

If the hyperparameter g is treated as a fixed but unknown quantity, its value must be selected carefully. In, e.g., [2], [4], [63], the consequences of selecting various fixed choices of g have been analysed. In [4], [64], the hyperparameter g was also treated as a random variable and integrated out of the marginal likelihood, thus avoiding the selection of a particular value for it. For the prior distribution on g, a special case of the beta prime or inverted beta distribution with pdf

$$p(g|\delta) = \frac{\delta - r}{r}(1 + g)^{-\delta/r} \ , \qquad \delta > r \ . \tag{32}$$

was used. The hyperparameter δ should be selected in the interval $r < \delta \leq 2r$ [4]. Besides having some desirable analytical properties, $p(g|\delta)$ reduces to the Jeffreys' prior and the reference prior for a linear regression model when $\delta = r$ [65]. However, since this prior is improper, it can only be used when the prior probability of the null model is zero.

4) The non-linear Parameters: The elicitation of the prior distribution on the non-linear parameters $\boldsymbol{\phi}_k$ is hard to do in general. In this paper, we therefore treat the case of non-linear parameters with a uniform prior of the form

$$p(\boldsymbol{\phi}_k|\mathcal{M}_k) = W_k^{-1}\mathbb{I}_{\Phi_k}(\boldsymbol{\phi}_k) \tag{33}$$

where $\mathbb{I}_{\Phi_k}(\cdot)$ is the indicator function on the support Φ_k and W_k is the normalisation constant. This uniform prior is often used for the non-linear parameters of sinusoidal and chirped signal models.

5) The Models: For the prior on the model, we select a uniform prior of the form $p(\mathcal{M}_k) = K^{-1}\mathbb{I}_{\mathcal{K}}(k)$ where $\mathcal{K} = \{1, 2, \ldots, K\}$. For a finite number of models, however, it is easy to use a different prior in our framework through (7).

B. Bayesian Inference

So far, we have elicited our probability model consisting of the observation model in (29) and the prior distributions on the model parameters. These distributions constitute the integrand of the integral representation of the marginal likelihood in (4), and we now evaluate this integral. After some algebra, the integrand can be rewritten as

$$\begin{aligned}
&p(\boldsymbol{x}|\boldsymbol{\alpha}_k, \boldsymbol{\psi}, \boldsymbol{\phi}_k, \sigma^2, \mathcal{M}_k)p(\boldsymbol{\alpha}_k|g, \sigma^2, \boldsymbol{\phi}_k, \mathcal{M}_k) \\
&\quad \times p(\boldsymbol{\psi})p(\sigma^2)p(g|\delta)p(\boldsymbol{\phi}_k|\mathcal{M}_k) \\
&\propto \mathcal{N}_r(\boldsymbol{\alpha}; g\boldsymbol{m}_k/(1+g), g\sigma^2[\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Z}_k(\boldsymbol{\phi}_k)]^{-1}/(1+g)) \\
&\quad \times \mathcal{N}_r(\boldsymbol{\psi}; \boldsymbol{m}_N, \sigma^2[\boldsymbol{B}^H\boldsymbol{B}]^{-1})\text{Inv-}\mathcal{G}\left(\sigma^2; \frac{N-l_N}{r}, \frac{N\hat{\sigma}_k^2}{r}\right) \\
&\quad \times \frac{m_N(\boldsymbol{x})}{(1+g)^{l_k/r}}\left(\frac{\hat{\sigma}_N^2}{\hat{\sigma}_k^2}\right)^{(N-l_N)/r} p(g|\delta)p(\boldsymbol{\phi}_k|\mathcal{M}_k)
\end{aligned} \tag{34}$$

where Inv-$\mathcal{G}$ is the inverse gamma distribution. Moreover, we have defined

$$\boldsymbol{m}_k \triangleq [\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Z}_k(\boldsymbol{\phi}_k)]^{-1}\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{x} \tag{35}$$

$$\boldsymbol{m}_N \triangleq (\boldsymbol{B}^H\boldsymbol{B})^{-1}\boldsymbol{B}^H\boldsymbol{x} \tag{36}$$

$$\hat{\sigma}_k^2 \triangleq \frac{\boldsymbol{x}^H(\boldsymbol{I}_N - \boldsymbol{P}_{\boldsymbol{B}} - \frac{g}{1+g}\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{x}}{N} \tag{37}$$

$$m_N(\boldsymbol{x}) \triangleq \frac{\Gamma((N-l_N)/r)}{(N\pi\hat{\sigma}_N^2)^{(N-l_N)/r}|\boldsymbol{B}^H\boldsymbol{B}|^{1/r}} \tag{38}$$

where $\boldsymbol{P}_{\boldsymbol{B}}$ and $\boldsymbol{P}_{\boldsymbol{Z}}$ are the orthogonal projection matrices of $\boldsymbol{B}$ and $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$, respectively, and $\hat{\sigma}_k^2$ is asymptotically equal to the maximum likelihood (ML) estimate of the noise variance in the limit $\hat{\sigma}_{\text{ML}}^2 = \lim_{g\to\infty}\hat{\sigma}_k^2$. The estimate $\hat{\sigma}_N^2$ is the estimated noise variance of the null model, and it is defined as $\hat{\sigma}_k^2$ for $\boldsymbol{P}_{\boldsymbol{Z}} = 0$. Finally, $m_N(\boldsymbol{x})$ is the unnormalised marginal likelihood of the null model. The linear parameters and the noise variance is now easily integrated out of the marginal likelihood in (4). Doing this, we obtain that

$$p(\boldsymbol{x}|g, \boldsymbol{\phi}_k, \mathcal{M}_k) \propto \frac{m_N(\boldsymbol{x})}{(1+g)^{l_k/r}}\left(\frac{\hat{\sigma}_N^2}{\hat{\sigma}_k^2}\right)^{(N-l_N)/r} \tag{39}$$

$$= \frac{m_N(\boldsymbol{x})(1+g)^{(N-l_N-l_k)/r}}{(1+g[1-R_k^2(\boldsymbol{\phi}_k)])^{(N-l_N)/r}} \tag{40}$$

which we define as the unnormalised marginal likelihood $m_k(\boldsymbol{x}|g, \boldsymbol{\phi}_k)$ of model $\mathcal{M}_k$ given g and $\boldsymbol{\phi}_k$. Moreover,

$$R_k^2(\boldsymbol{\phi}_k) \triangleq \frac{\boldsymbol{x}^H\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{x}}{\boldsymbol{x}^H(\boldsymbol{I}_N - \boldsymbol{P}_{\boldsymbol{B}})\boldsymbol{x}} \tag{41}$$

resembles the coefficient of determination from classical linear regression analysis where it measures how well the data set fits the regression. Whereas the linear parameters and the noise variance were easily integrated out of the marginal likelihood, the hyperparameter g and the non-linear parameters $\boldsymbol{\phi}_k$ are not. In the next two sections, we therefore propose approximate ways of performing the integration over these parameters.

IV. Known System Matrix

In this section, we consider the case where there are either no non-linear parameters or they are known.

A. Fixed Choices of g

We first assume that g is a fixed quantity. From (6) and (39), the Bayes' factor is therefore

$$\text{BF}[\mathcal{M}_k; \mathcal{M}_N|g, \boldsymbol{\phi}_k] = (1+g)^{-l_k/r}\left(\frac{\hat{\sigma}_N^2}{\hat{\sigma}_k^2}\right)^{(N-l_N)/r}. \tag{42}$$

With a uniform prior on the models, it follows from (7) that the Bayes' factor is proportional to the posterior distribution with pdf $p(\mathcal{M}_k|\boldsymbol{x}, g, \boldsymbol{\phi}_k)$ on the models. The model with the highest posterior probability is the solution to

$$\hat{k} = \arg\max_{k\in\mathcal{K}} p(\mathcal{M}_k|\boldsymbol{x}, g, \boldsymbol{\phi}_k) \tag{43}$$

$$= \arg\max_{k\in\mathcal{K}} \left[-(N-l_N)\ln\hat{\sigma}_k^2 - l_k\ln(1+g)\right]. \tag{44}$$

As alluded to in Sec. III-A3, the value of g is vital in model selection. From (40), we see that if $g \to \infty$, the Bayes' factor in (42) goes to zero. The null model is therefore always the most probable model, regardless of the information in the data (Bartlett's paradox). Another problem occurs if we assume that the least squares estimate $\boldsymbol{m}_k \to \infty$ or, equivalently, that $R_k^2(\boldsymbol{\phi}_k) \to 1$ so that the null model cannot be true. Although we would expect that the Bayes' factor would also go to infinity, it converges to the constant $(1+g)^{(N-l_n-l_k)/r}$, and this is called the information paradox [4], [35], [66]. For these two reasons, the value of g should depend on the data in some way. A local empirical Bayesian (EB) estimate is a data-dependent estimate of g, and it is the maximum of the marginal likelihood w.r.t. g [4]

$$g_k^{\text{EB}} = \arg\max_{g\in\mathbb{R}^+} p(\boldsymbol{x}|g, \boldsymbol{\phi}_k, \mathcal{M}_k) \tag{45}$$

$$= \max\left(\frac{(N-l_N)R_k^2(\boldsymbol{\phi}_k) - l_k}{(1-R_k^2(\boldsymbol{\phi}_k))l_k}, 0\right), \; l_k > 0 \tag{46}$$

where $\mathbb{R}^+$ is the set of non-negative real-valued numbers. This choice of g clearly resolves the information paradox. Inserting

the EB estimate of g into (44) gives for $R_k^2(\boldsymbol{\phi}_k) > l_k/N$ the empirical BIC (e-BIC)

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \big[-(N-l_N)\ln\hat{\sigma}^2_{\text{ML}} - l_k\ln(1+g_k^{\text{EB}}) + (N-l_N)\ln(1-l_k/(N-l_N))\big] \quad (47)$$

whose form is similar to most of the information criteria. When the null model is the noise-only model so that $l_N = 0$, these information criteria can be written as [20][6]

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \big[-2N\ln\hat{\sigma}^2_{\text{ML}} - r\nu_k h(\nu_k, N)\big] \quad (48)$$

where ν_k is the number of real-valued independent parameters in the model, and $h(\nu_k, N)$ is a penalty coefficient. For $h(\nu_k, N) = \{2, \ln N\}$, we get the AIC and the MDL, respectively. Note that ν_k is not always the same as the number of unknown parameters [30]. Moreover, if the penalty coefficient does not depend on the candidate model, ν_k may be interpreted as the number of independent parameters which are not in all candidate models. In nested models with white Gaussian noise and a known system matrix, this means that the noise variance parameter does not have to be counted as an independent parameter. Thus, selecting ν_k as either $\nu_k = 2l_k/r + 1$ or $\nu_k = 2l_k/r$ does not change, e.g., the AIC and the MDL.

1) Interpretation of the e-BIC: To gain some insight into the behaviour of the e-BIC, we here compare it to the AIC and the MDL in the context of a linear regression model with $N \gg l_k$ and $l_N = 0$. Under these assumptions, the penalty coefficient of the e-BIC in (47) reduces to

$$h(\nu_k, N) = \ln(1+g_k^{\text{EB}}) - N\ln(1-l_k/N)/l_k \quad (49)$$
$$\approx \ln(1+g_k^{\text{EB}}) + 1 = \ln(1+N\eta_k^{\text{EB}}/l_k) + 1 \quad (50)$$

where the approximation follows from the assumption that $N \gg l_k$ so that $\ln(1-l_k/N) \approx -l_k/N$. From the approximate e-BIC (ae-BIC) in (50), several interesting observations can be made. When the SNR is large enough to justify that $N\eta_k^{\text{EB}} \gg l_k$, the e-BIC is basically a corrected MDL 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 MDL tend to overestimate the model order [67]. When the estimated SNR on the other hand becomes so low that $N\eta_k^{\text{EB}} \ll l_k$, the e-BIC reduces to an AIC-like rule which has a constant penalty coefficient. In the extreme case of an estimated SNR of zero, the e-BIC reduces to the so-called no-name rule [20]. Interestingly, empirical studies [40], [68] have shown that the AIC performs better than the MDL 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 have demonstrated for the polynomial model in [62].

B. Integration over g

Another way to resolve the information paradox is by treating g as a random variable and integrate it out of the marginal likelihood. For the prior distribution on g in (32) and the unnormalised marginal likelihood in (40), we obtain the Bayes' factor given by

$$\text{BF}[\mathcal{M}_k;\mathcal{M}_N|\boldsymbol{\phi}_k] = \int_0^\infty \frac{m_k(\boldsymbol{x}|g,\boldsymbol{\phi}_k)}{m_N(\boldsymbol{x})} p(g|\delta)dg = \frac{\delta-r}{l_k+\delta-r}\,{}_2F_1\left(\frac{N-l_N}{r}, 1; \frac{l_k+\delta}{r}; R_k^2(\boldsymbol{\phi}_k)\right) \quad (51)$$

where ${}_2F_1$ is the Gaussian hypergeometric function [69, p. 314]. When N is large or $R_k^2(\boldsymbol{\phi}_k)$ is very close to one, numerical and computational problems with the evaluation of the Gaussian hypergeometric function may be encountered [70]. From a computational point of view, it may therefore not be advantageous to marginalise (51) w.r.t. g analytically. Instead, the Laplace approximation can be used as a simple alternative. Using the procedure outlined in Sec. II-B1 and the results in App. B, we get that

$$\text{BF}[\mathcal{M}_k;\mathcal{M}_N|\boldsymbol{\phi}_k] \approx \text{BF}[\mathcal{M}_k;\mathcal{M}_N|\hat{g},\boldsymbol{\phi}_k] \frac{\hat{g}(\delta-r)}{r(1+\hat{g})^{\delta/r}} \sqrt{2\pi\gamma(\hat{g}|\boldsymbol{\phi}_k)} \quad (52)$$

where $\hat{g} = \exp(\hat{\tau})$ and $\gamma(\hat{g}|\boldsymbol{\phi}_k)$ can be found from (83) and (84), respectively, with $v = 1$, $w = (N - l_N - l_k - \delta)/r$, and $u = (N-l_N)/r$. Since the marginal posterior distribution on g does not have a symmetric pdf and in order to avoid edge effects near $g = 0$, the Laplace approximation was made for the parametrisation $\tau = \ln g$ [4]. This parametrisation suggests that the posterior distribution on g is approximately a log-normal distribution. The model with the highest posterior probability can be found by maximising (52) w.r.t. the model index and this yields the Laplace BIC (lp-BIC)

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \big[-(N-l_N)\ln\hat{\sigma}_k^2 - l_k\ln(1+\hat{g}) - \delta\ln(1+\hat{g}) + r\ln\hat{g} + (r/2)\ln\gamma(\hat{g}|\boldsymbol{\phi}_k)\big] \quad (53)$$

Compared to the maximisation in (44), (53) differs in terms of the estimate of g and the last three terms. These terms account for the uncertainty in our point estimate of g. In Table I, we have compared the proposed model selection and comparison rules with the AIC, the MDL, and the MAP rule for regression models with a known system matrix and $l_N = 0$.

Rule	$h(\nu_k, N)$	Bayes' factor
AIC	2	See [21]
MDL	$\ln N$	See [21]
MAP	Model dependent	See [21]
e-BIC	$\ln(1+g_k^{\text{EB}}) - N\ln(1-l_k/N)/l_k$	(42)
ae-BIC	$\ln(1+g_k^{\text{EB}}) + 1$	(42)
lp-BIC	No short expression. See (53) instead.	(52)

TABLE I
PENALTY TERMS AND BAYES' FACTORS FOR REGRESSION MODELS WITH A KNOWN SYSTEM MATRIX AND THE NOISE-ONLY MODEL AS THE NULL MODEL.

V. UNKOWN NON-LINEAR PARAMETERS

In this section, the ρ_k real-valued and non-linear parameters $\boldsymbol{\phi}_k$ are also assumed unknown, and they must therefore be integrated out of the marginal likelihood in (40). Since an

[6]The cost function must be divided by $2r$ when the information criteria are used for model averaging and comparison in the so-called multi-modal approach [21].

analytical marginalisation is usually not possible, we here consider doing the joint integral over $\boldsymbol{\phi}_k$ and g using the Laplace approximation with the change of variables $\tau = \ln g$. Dividing (40) by $m_N(\boldsymbol{x})$ yields the following integral representation of the Bayes' factor in (6)

$$\text{BF}[\mathcal{M}_k;\mathcal{M}_N] = \int_{-\infty}^{\infty}\int_{\Phi_k} q(\boldsymbol{\phi}_k,\tau)d\boldsymbol{\phi}_k d\tau \tag{54}$$

where the integrand is given by

$$q(\boldsymbol{\phi}_k,\tau) = \text{BF}[\mathcal{M}_k;\mathcal{M}_N|\exp(\tau),\boldsymbol{\phi}_k]p(\boldsymbol{\phi}_k|\mathcal{M}_k)p(\tau|\delta) \tag{55}$$

with

$$p(\tau|\delta) = \exp(\tau)p(g|\delta)\big|_{g=\exp(\tau)} \ . \tag{56}$$

According to the procedure outlined in Sec. II-B1, we need to find the mode and Hessian of $\ln q(\boldsymbol{\phi}_k,\tau)$ to approximate the integrand by a normal pdf. For the uniform prior on $\boldsymbol{\phi}_k$ in (33), the mode of $\ln q(\boldsymbol{\phi}_k,\tau)$ w.r.t. $\boldsymbol{\phi}_k$ is given by

$$\begin{aligned}\hat{\boldsymbol{\phi}}_k^{\text{MAP}} &= \arg\max_{\boldsymbol{\phi}_k\in\Phi_k} p(\boldsymbol{x}|g,\boldsymbol{\phi}_k,\mathcal{M}_k) = \arg\max_{\boldsymbol{\phi}_k\in\Phi_k} p(\boldsymbol{x}|\boldsymbol{\phi}_k,\mathcal{M}_k)\\ &= \arg\max_{\boldsymbol{\phi}_k\in\Phi_k} R_k^2(\boldsymbol{\phi}_k) = \arg\max_{\boldsymbol{\phi}_k\in\Phi_k} C_k(\boldsymbol{\phi}_k)\end{aligned} \tag{57}$$

where we have defined

$$C_k(\boldsymbol{\phi}_k) \triangleq \boldsymbol{x}^H\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{x} \ . \tag{58}$$

Note that $C_k(\boldsymbol{\phi}_k)$ does not depend on the hyperparameter g (and equivalently τ) so the MAP estimator $\hat{\boldsymbol{\phi}}_k^{\text{MAP}}$ is independent of the prior on g. Depending on the structure of $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$, it might be hard to perform the maximisation of $C_k(\boldsymbol{\phi}_k)$. In App. C, we have therefore derived the first and second order differentials of an orthogonal projection matrix as these are useful in numerical optimisation algorithms for maximising $C_k(\boldsymbol{\phi}_k)$. We also note in passing that the MAP estimator is identical to the ML estimator for the non-linear regression model in (25). Evaluated at the mode $\hat{\boldsymbol{\phi}}_k^{\text{MAP}}$, the Hessian matrix $\boldsymbol{H}(\boldsymbol{\phi}_k)$ is given by

$$\boldsymbol{H}(\hat{\boldsymbol{\phi}}_k^{\text{MAP}}) = \frac{\exp(\tau)(N-l_N)}{rN[1+\exp(\tau)]\hat{\sigma}_k^2}\boldsymbol{D} \tag{59}$$

where we have defined

$$\boldsymbol{D} \triangleq \frac{\partial^2 C_k(\boldsymbol{\phi}_k)}{\partial\boldsymbol{\phi}_k\partial\boldsymbol{\phi}_k^T}\bigg|_{\boldsymbol{\phi}_k=\hat{\boldsymbol{\phi}}_k^{\text{MAP}}} \ . \tag{60}$$

Using the results in App. C, the (n,m)'th element of $\boldsymbol{D}$ can be written as

$$\begin{aligned}[\boldsymbol{D}]_{nm} = 2\text{Re}\Big\{&\boldsymbol{w}^H\big[\boldsymbol{\Lambda}_{nm} - \boldsymbol{T}_n\boldsymbol{S}_k^{-1}\boldsymbol{Z}_k^H(\hat{\boldsymbol{\phi}}_k^{\text{MAP}})\boldsymbol{T}_m\\ &- \boldsymbol{T}_m\boldsymbol{S}_k^{-1}\boldsymbol{Z}_k^H(\hat{\boldsymbol{\phi}}_k^{\text{MAP}})\boldsymbol{T}_n\big]\boldsymbol{m}_k + \boldsymbol{w}^H\boldsymbol{T}_n\boldsymbol{S}_k^{-1}\boldsymbol{T}_m^H\boldsymbol{w}\\ &- \boldsymbol{m}_k^H\boldsymbol{T}_n^H(\boldsymbol{I}_N - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{T}_m\boldsymbol{m}_k\Big\}\end{aligned} \tag{61}$$

where we have defined

$$\boldsymbol{w} \triangleq \boldsymbol{x} - \boldsymbol{Z}_k(\hat{\boldsymbol{\phi}}_k^{\text{MAP}})\boldsymbol{m}_k \tag{62}$$

$$\boldsymbol{S}_k \triangleq \boldsymbol{Z}_k^H(\hat{\boldsymbol{\phi}}_k^{\text{MAP}})\boldsymbol{Z}_k(\hat{\boldsymbol{\phi}}_k^{\text{MAP}}) \tag{63}$$

$$\boldsymbol{T}_i \triangleq \frac{\partial\boldsymbol{Z}_k(\boldsymbol{\phi}_k)}{\partial\phi_i}\bigg|_{\boldsymbol{\phi}_k=\hat{\boldsymbol{\phi}}_k^{\text{MAP}}} \tag{64}$$

$$\boldsymbol{\Lambda}_{nm} \triangleq \frac{\partial^2\boldsymbol{Z}_k(\boldsymbol{\phi}_k)}{\partial\phi_n\partial\phi_m}\bigg|_{\boldsymbol{\phi}_k=\hat{\boldsymbol{\phi}}_k^{\text{MAP}}} \ . \tag{65}$$

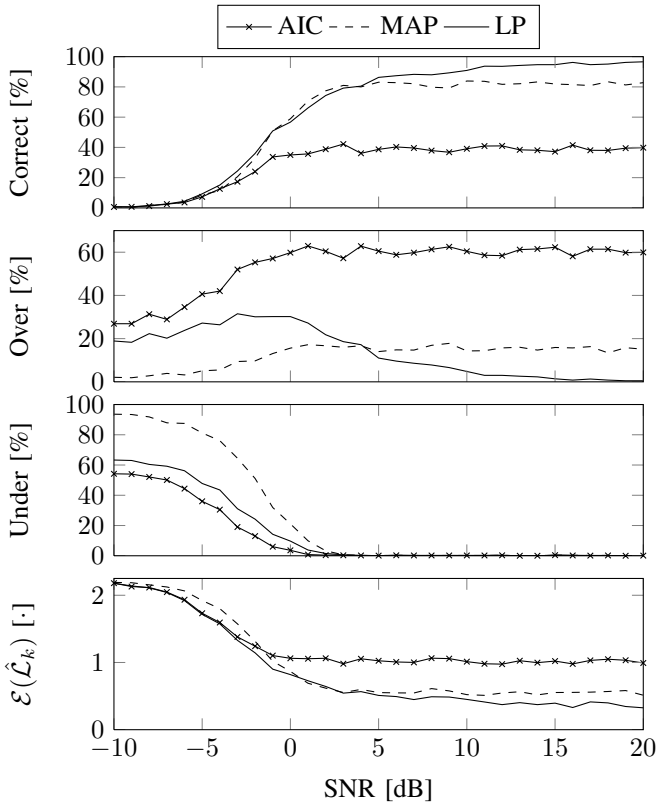

Fig. 2. The first three plots show the percentage of correctly detected, overestimated, and underestimated number of harmonic components versus the SNR for the harmonic signal model. The last plot shows the RMSE of the estimated number of harmonic components.

As we demonstrate in the Sec. VI, the value of $[\boldsymbol{D}]_{nm}$ can often be approximated by only the last term in (61).

Since $\hat{\boldsymbol{\phi}}_k^{\text{MAP}}$ does not depend on the value of τ, the mode and second-order derivative of $\ln q(\boldsymbol{\phi}_k,\tau)$ w.r.t. τ are therefore the same as in Sec. IV-B and can be found in App. B with $v=1$, $w=(N-l_N-l_k-\delta)/r$, and $u=(N-l_N)/r$. Thus, the Laplace approximation of the Bayes' factor in (54) is

$$\begin{aligned}\text{BF}[\mathcal{M}_k;\mathcal{M}_N] \approx \text{BF}[\mathcal{M}_k;\mathcal{M}_N|\hat{g},\hat{\boldsymbol{\phi}}_k^{\text{MAP}}]\frac{\hat{g}(\delta-r)}{r(1+\hat{g})^{\delta/r}}\\ \times W_k^{-1}(2\pi)^{(\rho_k+1)/2}\sqrt{\gamma(\hat{g}|\hat{\boldsymbol{\phi}}_k^{\text{MAP}})|-\boldsymbol{H}(\hat{\boldsymbol{\phi}}_k^{\text{MAP}})|^{-1/2}} \ .\end{aligned} \tag{66}$$

When $q(\boldsymbol{\phi}_k,\tau)$ consists of multiple, significant, and well-separated peaks, the integral in (54) can be approximated by a Laplace approximation to each peak at their respective modes [55]. In this case, the Bayes' factor in (66) will be a sum over each of these peaks. Since it is not obvious how the number of peaks should be selected in a computationally simple manner, we consider only one peak in the simulations in Sec. VI. Although this is often a crude approximation for low SNRs, we demonstrate that other model selection rules are still outperformed.

VI. Simulations

We demonstrate the applicability of our model comparison algorithms by three simulation examples. In the first example,

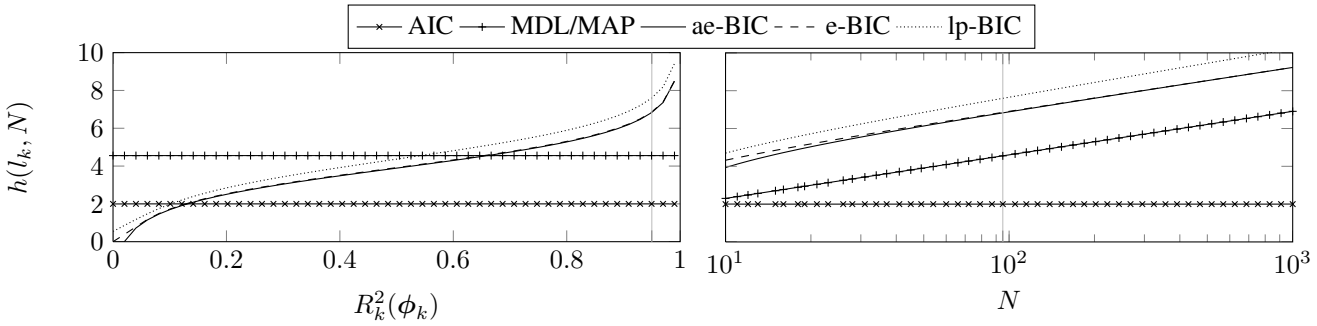

Fig. 1. Interpretation of the various information criteria for $l_k = 5$. The plots show the penalty coefficient $h(l_k, N)$ as a function of $R_k^2(\boldsymbol{\phi}_k)$ and the number of data points N. In the left plot, $N = 95$, and in the right, $R_k^2(\boldsymbol{\phi}_k) = 0.95$ for the e-BIC, the ae-BIC, and the lp-BIC.

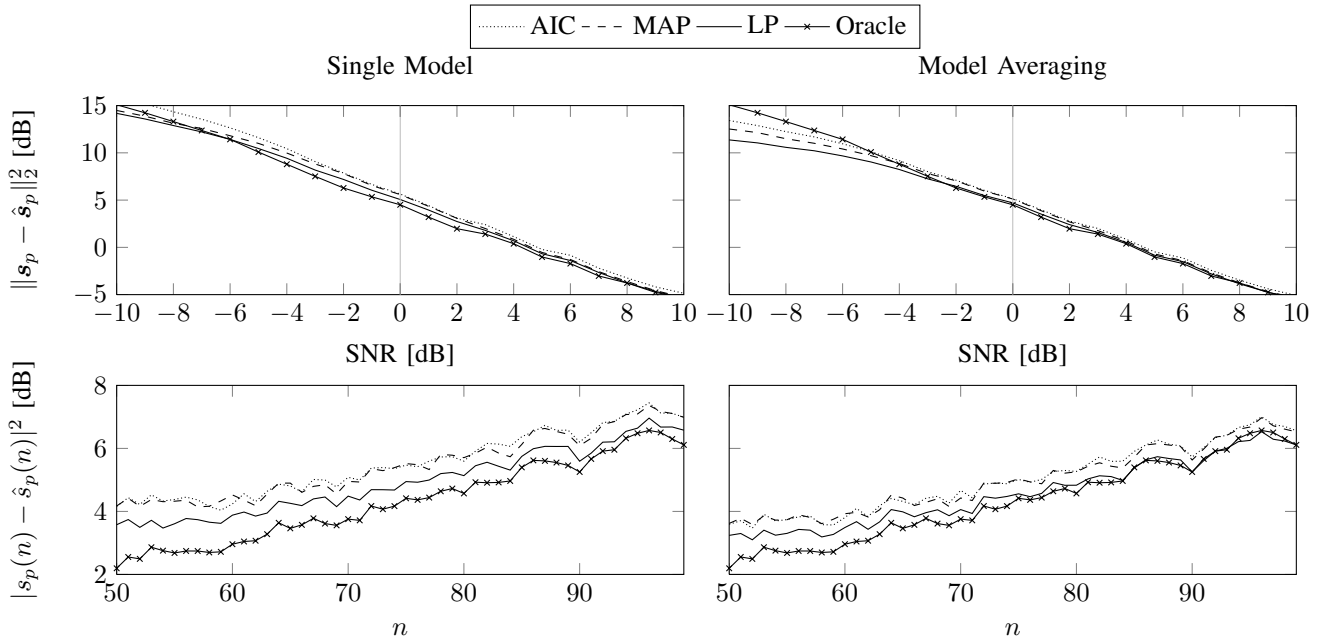

Fig. 3. Prediction performance versus the SNR (top row) and versus the prediction step at an SNR of 0 dB (bottom row) for a periodic signal model. In the plots in the left column, only the model with the highest posterior probability was used whereas all models were used in the plots in the right column.

we compare the penalty coefficient of our proposed algorithms with the penalty coefficient of the AIC, the MDL, and the MAP rule. In the second and third example, we consider model comparison in models containing unknown nonlinear parameters. Specifically, we first consider a periodic signal model which consists of a single non-linear parameter, the fundamental frequency, and then consider the uniform linear array model which consists of multiple non-linear parameters, the direction of arrivals (DOA). Similar simulations comparing the performance of the e-BIC in (42) and the lp-BIC in (52) to other order selection rules for linear and polynomial models can be found in [4] and [62], respectively. The simulation code can be found at http://kom.aau.dk/~jkn/publications/publications.php.

A. Penalty Coefficient

In Sec. IV-A1, we considered the interpretation of the AIC [23] and the MDL [26], [27] for a regression model with a known system matrix when the null model is the noise-only model and $N \gg l_k$. For the linear regression model, the MDL and the MAP are equivalent [40]. Here, we give some more insight by use of a simple simulation example in which the penalty coefficients of the AIC, the MDL/MAP, the e-BIC, the approximate e-BIC (ae-BIC), and the lp-BIC methods were found as a function of the coefficient of determination $R_k^2(\boldsymbol{\phi}_k)$ and the number of data points N. We fixed the number of linear parameters to $l_k = 5$, and Fig. 1 shows the results. In the left plot, the penalty coefficients $h(l_k, N)$ were computed as a function of $R_k^2(\boldsymbol{\phi}_k)$ for $N = 95$. Since the AIC and the MDL/MAP do not depend on the data, their penalty coefficients were constant. On the other hand, the penalty coefficients of the e-BIC, the ae-BIC, and the lp-BIC are data dependent and increased with the coefficient of determination. In the right plot, the penalty coefficients $h(l_k, N)$ were computed as a function of the number of data points N for $R_k^2(\boldsymbol{\phi}_k) = 0.95$. Note that the MDL/MAP had the same trend as the e-BIC, the ae-BIC, and the lp-BIC although shifted. The vertical distance between these penalties depends on the particular value of $R_k^2(\boldsymbol{\phi}_k)$. In Fig. 1, we set $R_k^2(\boldsymbol{\phi}_k) = 0.95$, but if $R_k^2(\boldsymbol{\phi}_k) \approx 0.648$ was selected instead, the e-BIC and the MDL/MAP would coincide for large values

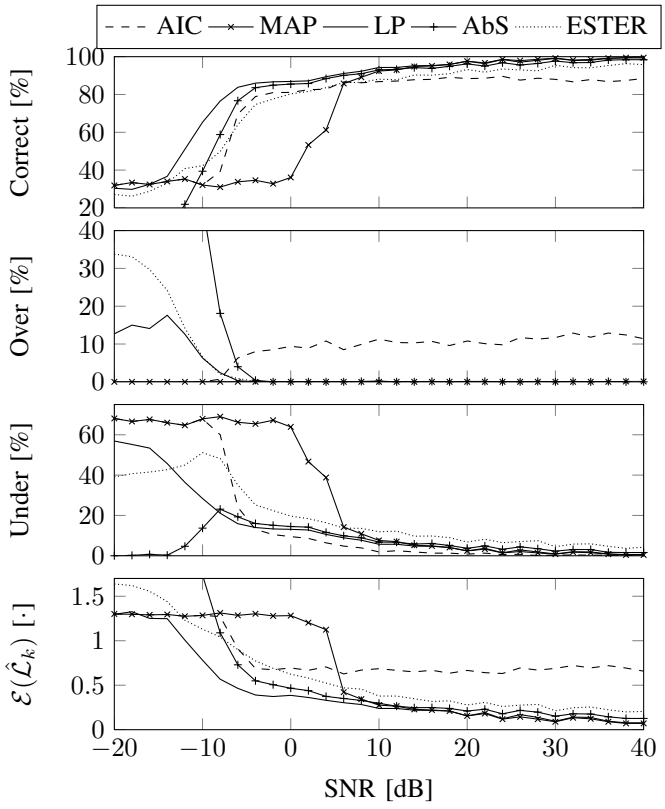

Fig. 4. The first three plots show the percentage of correctly detected, overestimated, and underestimated number of sources versus the SNR for the uniform linear array model. The last plot shows the RMSE of the estimated number of sources.

of N.

B. Periodic Signal Model

We consider a complex periodic signal model given by

$$\mathcal{M}_k: \quad x(n) = \sum_{i=1}^{L} \alpha_i \exp(j\omega i n)\mathbb{I}_{\mathcal{L}_k}(i) + e(n) \tag{67}$$

for $n = 0, 1, \ldots, N-1$ where $\mathbb{I}_{\mathcal{L}_k}(i)$ indicates whether the i'th harmonic component is included in the model $\mathcal{M}_k$ or not. This model is a special case of the model in (25) with the null model being the noise-only model, $\phi_k = \omega$, and $\boldsymbol{\alpha}_k$ being the complex amplitudes. Since no closed-form solution exists for the posterior distribution on the models for the periodic signal model, we consider the approximation in (66) which we refer to as the Laplace (LP) method. The method is compared to the AIC and the asymptotic MAP rule by Djuric with the latter having the penalty coefficient in (48) given by [9]

$$h(l_k, N) = \ln N + \frac{3}{2l_k} \ln N \ . \tag{68}$$

For the periodic signal model, the Hessian matrix in (59) is a scalar which can be approximated by [71]

$$H(\hat{\omega}_k^{\text{MAP}}) \approx -\frac{\hat{g}N(N^2-1)\sum_{i=1}^{L} |[\boldsymbol{m}_k]_i|^2 i^2 \mathbb{I}_{\mathcal{L}_k}(i)}{6(1+\hat{g})\hat{\sigma}_k^2} \tag{69}$$

where $\hat{\omega}_k^{\text{MAP}}$ is the ML estimate of the fundamental frequency. In the simulations, we set the maximum number of harmonic components to $L = 10$ and considered $K = 2^L - 1 = 1023$ models. Zero prior probability was assigned to the noise-only model as the model comparison performance was evaluated against the SNR. Moreover, this permits the use of the improper prior $p(g|\delta = r = 1)$ since g is now a common parameter in all models. For each SNR from -10 dB to 20 dB in steps of 1 dB, we ran 1000 Monte Carlo runs. As recommended in [21], a data vector $\boldsymbol{x}$ consisting of $N = 50$ samples was generated in each run by first randomly selecting a model from a uniform prior on the models. For this model, we then randomly selected the fundamental frequency and the phases of the complex amplitudes from a uniform distribution on the interval $[(2\pi)^{-1}, 2\pi/\max(\mathcal{L}_k)]$ and $[0, 2\pi]$, respectively. The amplitudes of the harmonics in the selected model were all set to one. Finally a complex noise vector was generated and normalised so that the data had the desired SNR. Besides generating a data vector, we also generated a vector $\boldsymbol{x}_p$ of unobserved data for $n = N, N+1, \ldots 2N-1$.

In Fig. 2, the percentage of correctly detected models, overestimated models, underestimated models, and the root-mean-squared-error (RMSE) of the estimated model versus the SNR are shown. The RMSE is defined as

$$\mathcal{E}(\hat{\mathcal{L}}_k) = \sqrt{\sum_{i=1}^{L} (\mathbb{I}_{\mathcal{L}_k}(i) - \mathbb{I}_{\hat{\mathcal{L}}_k}(i))^2} \tag{70}$$

where $\hat{\mathcal{L}}_k$ is the set containing the harmonic numbers of the most likely model. For an SNR below 0 dB, the LP method and the asymptotic MAP rule had a similar performance and were better than the AIC. For SNRs above 5 dB, the LP method also outperformed the asymptotic MAP rule. In terms of the RMSE, similar observations are made except that the asymptotic MAP rule performs worse than the other methods for low SNRs. However, it should be noted that the percentage of correctly detected models is not necessarily the best way of benchmarking model selection methods. As exemplified in [21], the true model does not always give the best prediction performance, and it may therefore be advantageous to either over- or underestimate the model order. Using the same Monte Carlo setup as above, we have therefore also investigated the prediction performance, and the results are shown in Fig. 3. In the plots in the left column, only the single model with the largest posterior probability was used for making the predictions of the predictable part $\boldsymbol{s}_p$ whereas all models were used as in (8) in the plots in the right column. The prediction based on a single model and all models was the mean of $p(\boldsymbol{x}_p|\boldsymbol{x}, \mathcal{M}_k)$ and $p(\boldsymbol{x}_p|\boldsymbol{x})$, respectively, where the latter depends on the former as in (8) with

$$p(\boldsymbol{x}_p|\boldsymbol{x}, \mathcal{M}_k) = \int_{\Theta_k} p(\boldsymbol{x}_p|\boldsymbol{\theta}_k, \mathcal{M}_k) p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k) d\boldsymbol{\theta}_k \ . \tag{71}$$

In the top row, the MSE of the total prediction error versus the SNR is shown, and in the bottom row, the MSE of the prediction error for each prediction step at an SNR of 0 dB is shown. In the four plots, the Oracle knew the true model but not the model parameters. From the four figures, we see again that the LP method outperformed the other methods with the AIC being the overall worst. For low SNRs, we also see that the MSE of the prediction errors were significantly lower

when model averaging was used. Moreover, we see that the performance was also better than the Oracle's performance and this demonstrates, as discussed above, that the true model does not always give the best prediction performance. For high SNRs, only the AIC performed slightly worse than the other methods which performed almost as well as the Oracle. Moreover, there was basically no difference between the single and multi-model predictions since a single model received all posterior probability.

C. Uniform Linear Array Signal Model

In the third and final simulation example, we consider the problem of estimation the number of narrowband source signals impinging on a uniform linear array (ULA) consisting of M calibrated sensors. For this problem, the model for the m'th sensor signal is given by [22, Ch. 6]

$$\mathcal{M}_k: \quad y_m(n) = \sum_{i=1}^{l_k} s_i(n)\exp(-j\omega_i n) + e_m(n) \qquad (72)$$

for $n = 0, 1, \ldots, N-1$ where ω_i is the spatial frequency in radians per sample of the i'th source. The spatial frequency is related to the direction of arrival (DOA) θ_i of the source signal by

$$\omega_i = \frac{\omega_c d \sin(\theta_i)}{v} \qquad (73)$$

where ω_c, d, and v are the carrier frequency in radians per second, the sensor distance in meters, and the propagation speed in meters per second, respectively. The signal $s_i(n)$ is the baseband signal of the i'th source. The signal model in (72) can be written into the form of (25) as

$$\text{vec}(\boldsymbol{Y}) = (\boldsymbol{I}_N \otimes \boldsymbol{Z}(\boldsymbol{\omega}_k))\text{vec}(\boldsymbol{S}_k) + \text{vec}(\boldsymbol{E}) \qquad (74)$$

where $\text{vec}(\cdot)$ and $\otimes$ denote the vectorisation and the Kronecker product, respectively. The $M \times N$ matrices $\boldsymbol{Y}$ and $\boldsymbol{E}$ contain the observed sensor signals, and the noise realisations, and the $l_k \times N$ matrix $\boldsymbol{S}_k$ contains the baseband signals. Finally, the $M \times l_k$ matrix $\boldsymbol{Z}_k(\boldsymbol{\omega})$ contains the l_k steering vectors with the (m, i)'th element being given by $\exp(-j\omega_i(m-1))$. As in the previous example, no closed-form expression exists for the posterior distribution on the models, and we therefore again consider the Laplace approximation in (66). By only keeping the last term of (61) and by making the approximation $\boldsymbol{Z}(\boldsymbol{\omega}_k)^H\boldsymbol{Z}(\boldsymbol{\omega}_k) \approx M\boldsymbol{I}_{l_k}$, the determinant of the negative of the Hessian matrix in (59) can be approximated by

$$|-\boldsymbol{H}(\hat{\boldsymbol{\omega}}_k^{\text{MAP}})| \approx \left(\frac{\hat{g}M^3}{6(1+\hat{g})\hat{\sigma}_k^2}\right)^{l_k}\prod_{i=1}^{l_k}\sum_{n=0}^{N-1}|s_i(n)|^2 \qquad (75)$$

where $\hat{\boldsymbol{\omega}}_k^{\text{MAP}}$ coincides with the maximum likelihood estimate of $\boldsymbol{\omega}$ which we have computed using the RELAX algorithm [72]. Using a Monte Carlo simulation consisting of 1000 runs for every SNR from -20 dB to 40 dB in steps of 2 dB, we evaluated the model detection performance for $N = 100$ snapshots and $M = 10$ sensors. As in the previous simulation, we generated the model parameters at random in every run with the baseband signals being realisations from a complex-valued white Gaussian process. The true number of sources was either one, two, or three. In addition to comparing the proposed method to the AIC and the asymptotic MAP rule, we also compared to two subspace-based methods which are often used in array processing. These are the MUSIC method using angle between subspaces (AbS) [33], [73] and the estimation error (ESTER) method [31] based on ESPRIT. Since neither of these methods are able to detect whether a source is present or not, the all-noise model was not included in the set of candidate models which was set to consist of maximum $K = 5$ sources. Fig. 4 shows the results of the simulation. The proposed method (LP) performed better than the other rules for SNRs up to approximately 15 dB where the asymptotic MAP rule achieved the same performance. For low SNRs, the AIC performed better than the asymptotic MAP rule. The MUSIC and ESTER methods performed well across all SNRs and only slightly worse than the proposed method. All methods except the AIC seem to be consistent order selection rules.

VII. Conclusion

Model comparison and selection is a difficult and important problem and a lot of methods have therefore been proposed. In this paper, we first gave an overview over how model comparison is performed for any model in a Bayesian framework. We also discussed the two major issues of doing the model comparison in a Bayesian framework, namely the elicitation of prior distributions and the evaluation of the marginal likelihood. Specifically, we reviewed the conditions for using improper prior distributions, and we briefly discussed approximate numerical and analytical algorithms for evaluating the marginal likelihood. In the second part of the paper, we analysed a general regression model in a Bayesian framework. The model consisted of both linear and non-linear parameters, and we used and motivated a prior of the same form as the Zellner's g-prior for this model. Many of the information criteria can be interpreted in a new light using this model with known non-linear parameters. These interpretations also gave insight into why the AIC often overestimate the model complexity for a high SNR, and why the MDL underestimate the model complexity for a low SNR. For unknown non-linear parameters, we proposed an approximate way of integrating them out of the marginal likelihood using the Laplace approximation, and we demonstrated through two simulation examples that our proposed model comparison and selection algorithm outperformed other algorithms such as the AIC, the MDL, and the asymptotic MAP rule both in terms of detecting the true model and in making predictions.

Appendix A
Fisher Information Matrix for the Observation Model

Let $\boldsymbol{\gamma}$ denote a mixed parameter vector of complex-valued and real-valued parameters. Using the procedure in [74, App. 15C], it can be shown that the (n, m)'th element of the Fisher information matrix (FIM) for the normal distribution

$\mathcal{N}_r(\boldsymbol{x};\boldsymbol{\mu}(\boldsymbol{\gamma});\boldsymbol{\Sigma}(\boldsymbol{\gamma}))$ is given by

$$[\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})]_{nm} = \frac{1}{r}\left(\frac{\partial\boldsymbol{\mu}^*(\boldsymbol{\gamma})}{\partial\gamma_n^*}\right)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma})\left(\frac{\partial\boldsymbol{\mu}^*(\boldsymbol{\gamma})}{\partial\gamma_m^*}\right)^* + \frac{1}{r}\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\gamma})}{\partial\gamma_n^*}\right)^H\boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma})\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\gamma})}{\partial\gamma_m^*}\right) + \frac{1}{r}\text{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma})\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\gamma})}{\partial\gamma_n^*}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma})\left(\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\gamma})}{\partial\gamma_m^*}\right)^H\right). \quad (76)$$

For the observation model in (29), the parameter vector is given by $\boldsymbol{\gamma} = \begin{bmatrix}\boldsymbol{\psi}^T & \boldsymbol{\alpha}_k^T & \boldsymbol{\phi}_k^T & \sigma^2\end{bmatrix}^T$, and the mean vector and covariance matrix are given by

$$\boldsymbol{\mu}(\boldsymbol{\gamma}) = \boldsymbol{B}\boldsymbol{\psi} + \boldsymbol{Z}_k(\boldsymbol{\phi}_k)\boldsymbol{\alpha}_k \quad (77)$$

$$\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \sigma^2\boldsymbol{I}_N\ . \quad (78)$$

Computing the derivatives in (76) for the observation model in (29) yields the FIM given by

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma}) = \frac{1}{\sigma^2}\begin{bmatrix}\boldsymbol{B}^H\boldsymbol{B} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\mathcal{I}}(\boldsymbol{\alpha}_k,\boldsymbol{\phi}_k) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \frac{N}{r\sigma^2}\end{bmatrix} \quad (79)$$

where

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\alpha}_k,\boldsymbol{\phi}_k) = \begin{bmatrix}\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Z}_k(\boldsymbol{\phi}_k) & \boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Q}_k(\boldsymbol{\phi}_k)\\ \boldsymbol{Q}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Z}_k(\boldsymbol{\phi}_k) & \frac{2}{r}\text{Re}\left(\boldsymbol{Q}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Q}_k(\boldsymbol{\phi}_k)\right)\end{bmatrix}$$

$$\boldsymbol{Q}_k(\boldsymbol{\phi}_k) \triangleq \frac{\partial(\boldsymbol{Z}_k(\boldsymbol{\phi}_k)\boldsymbol{\alpha}_k)}{\partial\boldsymbol{\phi}_k}\ .$$

Note that $\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})$ is block diagonal which follows from the assumption that $\boldsymbol{B}^H\boldsymbol{Z}_k(\boldsymbol{\phi}_k) = \boldsymbol{0}$.

APPENDIX B
LAPLACE APPROXIMATION WITH THE HYPER-G PRIOR

For the hyper-g prior in (32), the integral in (51) with the change of variables to $\tau = \ln g$ can be written in the form

$$\int_0^\infty g^{v-1}(1+g)^w[1+g(1-R_k^2(\boldsymbol{\phi}_k))]^{-u}dg = \int_{-\infty}^\infty \exp(v\tau)(1+\exp(\tau))^w[1+\exp(\tau)(1-R_k^2(\boldsymbol{\phi}_k))]^{-u}d\tau\ .$$

Taking the derivative of the logarithm of the integrand and equating to zero lead to the quadratic equation

$$0 = \alpha_\tau\exp(2\tau) + \beta_\tau\exp(\tau) + v \quad (80)$$

where we have defined

$$\alpha_\tau \triangleq (1-R_k^2(\boldsymbol{\phi}_k))(v+w-u) \quad (81)$$

$$\beta_\tau \triangleq (u-v)R_k^2(\boldsymbol{\phi}_k) + 2v + w - u \quad (82)$$

For $u - w > v$, the only positive solution to this quadratic equation is

$$\hat{\tau} = \ln\left(\frac{\beta_\tau + \sqrt{\beta_\tau^2 - 4\alpha_\tau v}}{-2\alpha_\tau}\right) \quad (83)$$

which is the mode of the normal approximation to the integrand. The corresponding variance at this mode with $\hat{g} = \exp(\hat{\tau})$ is

$$\gamma(\hat{g}|\boldsymbol{\phi}_k) = \left[\frac{\hat{g}u(1-R_k^2(\boldsymbol{\phi}_k))}{[1+\hat{g}(1-R_k^2(\boldsymbol{\phi}_k))]^2} - \frac{\hat{g}w}{(1+\hat{g})^2}\right]^{-1}. \quad (84)$$

APPENDIX C
DIFFERENTIALS OF A PROJECTION MATRIX

Let $\boldsymbol{P} = \boldsymbol{G}(\boldsymbol{G}^H\boldsymbol{G})^{-1}\boldsymbol{G}^H$ denote an orthogonal projection matrix, and let $\boldsymbol{S} = \boldsymbol{G}^H\boldsymbol{G}$ denote an inner matrix product. The differential of $\boldsymbol{S}$ is then given by

$$d\boldsymbol{S} = (d\boldsymbol{G})^H\boldsymbol{G} + \boldsymbol{G}^H(d\boldsymbol{G})\ . \quad (85)$$

This result can be used to show that

$$d\boldsymbol{S}^{-1} = -\boldsymbol{S}^{-1}[(d\boldsymbol{G})^H\boldsymbol{G} + \boldsymbol{G}^H(d\boldsymbol{G})]\boldsymbol{S}^{-1}\ , \quad (86)$$

and that

$$d\boldsymbol{P} = \boldsymbol{P}^\perp(d\boldsymbol{G})\boldsymbol{S}^{-1}\boldsymbol{G}^H + \boldsymbol{G}\boldsymbol{S}^{-1}(d\boldsymbol{G})^H\boldsymbol{P}^\perp \quad (87)$$

where $\boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{P}$ is the complementary projection of $\boldsymbol{P}$. Let δ denote another differential operator. From the above results, we obtain after some algebra that

$$\begin{aligned}\delta(d\boldsymbol{P}) &= \boldsymbol{P}^\perp(\delta(d\boldsymbol{G}))\boldsymbol{S}^{-1}\boldsymbol{G}^H + \boldsymbol{G}\boldsymbol{S}^{-1}(\delta(d\boldsymbol{G}))^H\boldsymbol{P}^\perp\\ &+ \boldsymbol{P}^\perp\left[(d\boldsymbol{G})\boldsymbol{S}^{-1}(\delta\boldsymbol{G})^H + (\delta\boldsymbol{G})\boldsymbol{S}^{-1}(d\boldsymbol{G})^H\right]\boldsymbol{P}^\perp\\ &- \boldsymbol{P}^\perp\left[(\delta\boldsymbol{G})\boldsymbol{S}^{-1}\boldsymbol{G}^H(d\boldsymbol{G}) + (d\boldsymbol{G})\boldsymbol{S}^{-1}\boldsymbol{G}^H(\delta\boldsymbol{G})\right]\boldsymbol{S}^{-1}\boldsymbol{G}^H\\ &- \boldsymbol{G}\boldsymbol{S}^{-1}\left[(\delta\boldsymbol{G})^H\boldsymbol{G}\boldsymbol{S}^{-1}(d\boldsymbol{G})^H + (d\boldsymbol{G})^H\boldsymbol{G}\boldsymbol{S}^{-1}(\delta\boldsymbol{G})^H\right]\boldsymbol{P}^\perp\\ &- \boldsymbol{G}\boldsymbol{S}^{-1}\left[(d\boldsymbol{G})^H\boldsymbol{P}^\perp(\delta\boldsymbol{G}) + (\delta\boldsymbol{G})^H\boldsymbol{P}^\perp(d\boldsymbol{G})\right]\boldsymbol{S}^{-1}\boldsymbol{G}^H\ .\end{aligned}$$

REFERENCES

[1] G. E. P. Box and N. R. Draper, *Empirical model-building and response surface*. John Wiley & Sons, Inc., Jan. 1987.

[2] E. I. George and D. P. Foster, "Calibration and empirical Bayes variable selection," *Biometrika*, vol. 87, no. 4, pp. 731–747, Dec. 2000.

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

[4] 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.

[5] L. Kavalieris and E. J. Hannan, "Determining the number of terms in a trigonometric regression," *J. of Time Series Analysis*, vol. 15, no. 6, pp. 613–625, Nov. 1994.

[6] B. G. Quinn, "Estimating the number of terms in a sinusoidal regression," *J. of Time Series Analysis*, vol. 10, no. 1, pp. 71–75, Jan. 1989.

[7] 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.

[8] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2498–2517, Apr. 2006.

[9] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, B. H. Juang, Ed. Morgan & Claypool, 2009.

[10] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Royal Stat. Soc., Series B*, vol. 41, no. 2, pp. 190–195, 1979.

[11] G. Liang, D. M. Wilkes, and J. A. Cadzow, "Arma model order estimation based on the eigenvalues of the covariance matrix," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3003–3009, Oct. 1993.

[12] B. Choi, *ARMA model identification*. Springer-Verlag, Jun. 1992.

[13] S. Koreisha and G. Yoshimoto, "A comparison among identification procedures for autoregressive moving average models," *International Statistical Review*, vol. 59, no. 1, pp. 37–57, Apr. 1991.

[14] 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.

[15] 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.

[16] Z. Liang, R. Jaszczak, and R. Coleman, "Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing," *IEEE Trans. Nucl. Sci.*, vol. 39, no. 4, pp. 1126–1133, Aug. 1992.

[17] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Adv. in Neural Inf. Process. Syst.*, 2000, pp. 554–560.

[18] 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.

[19] C. R. Rao and Y. Wu, "On model selection," *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, vol. 38, pp. 1–57, 2001.

[20] 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, Jul. 2004.

[21] 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.

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

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

[24] 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.

[25] S. Konishi and G. Kitagawa, "Generalised information criteria in model selection," *Biometrika*, vol. 83, no. 4, pp. 875–890, Dec. 1996.

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

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

[28] ——, "Estimation of structure by minimum description length," *Circuits, Systems, and Signal Process.*, vol. 1, no. 3, pp. 395–406, 1982.

[29] ——, "A predictive least-squares principle," *IMA J. Math. Control Inf.*, vol. 3, no. 2–3, pp. 211–222, 1986.

[30] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 387–392, Apr. 1985.

[31] 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.

[32] 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, Jul. 2007.

[33] 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.

[34] J. O. Berger and L. R. Pericchi, "The intrinsic Bayes factor for model selection and prediction," *J. Amer. Statistical Assoc.*, vol. 91, no. 433, pp. 109–122, Mar. 1996.

[35] ——, "Objective Bayesian methods for model selection: Introduction and comparison," *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, vol. 38, pp. 135–207, 2001.

[36] L. Wasserman, "Bayesian model selection and model averaging," *J. of Mathematical Psychology*, vol. 44, no. 1, pp. 92–107, Mar. 2000.

[37] A. F. Deltell, "Objective bayes criteria for variable selection," Ph.D. dissertation, Universitat de València, 2011.

[38] P. M. Djuric and S. M. Kay, "Model selection based on Bayesian predictive densities and multiple data records," *IEEE Trans. Signal Process.*, vol. 42, no. 7, pp. 1685–1699, Jul. 1994.

[39] P. M. Djuric, "A model selection rule for sinusoids in white Gaussian noise," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, Jul. 1996.

[40] ——, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[41] 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, Sep. 2012.

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

[43] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Jun. 2002.

[44] R. Strachan and H. K. v. Dijk, "Improper priors with well defined Bayes' factors," Department of Economics, University of Leicester, Discussion Papers in Economics 05/4, 2005.

[45] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, May 2001.

[46] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer-Verlag New York, Inc., Jul. 2004.

[47] C. Andrieu, A. Doucet, and C. P. Robert, "Computational advances for and from Bayesian analysis," *Statist. Sci.*, vol. 19, no. 1, pp. 118–127, Feb. 2004.

[48] S. Chib, "Marginal likelihood from the Gibbs output," *J. Amer. Statistical Assoc.*, vol. 90, no. 432, pp. 1313–1321, Dec. 1995.

[49] S. Chib and I. Jeliazkov, "Marginal likelihood from the Metropolis-Hastings output," *J. Amer. Statistical Assoc.*, vol. 96, no. 453, pp. 270–281, Mar. 2001.

[50] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.

[51] M. Hong, M. F. Bugallo, and P. M. Djuric, "Joint model selection and parameter estimation by population Monte Carlo simulation," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 3, pp. 526–539, Jun. 2010.

[52] C. Han and B. P. Carlin, "Markov chain Monte Carlo methods for computing Bayes factors: A comparative review," *J. Amer. Statistical Assoc.*, vol. 96, no. 455, pp. 1122–1132, Sep. 2001.

[53] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, Aug. 2006.

[54] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal," *J. Amer. Statistical Assoc.*, vol. 81, no. 393, pp. 82–86, Mar. 1986.

[55] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Jul. 2003.

[56] A. Zellner and A. Siow, "Posterior odds ratios for selected regression hypotheses," *Trabajos de Estadística y de Investigación Operativa*, vol. 31, pp. 585–603, 1980.

[57] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. 4, no. 3, pp. 227–241, 1968.

[58] ——, *Probability Theory : The Logic of Science*, G. L. Bretthorst, Ed. Cambridge University Press, Apr. 2003.

[59] G. L. Bretthorst, "The near-irrelevance of sampling frequency distributions," in *Max. Entropy and Bayesian Methods*, 1999, pp. 21–46.

[60] 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. D. Reidel, Dordrecht-Holland, 1987, pp. 1–37.

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

[62] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "Bayesian model comparison and the BIC for regression models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 6362–6366.

[63] C. Fernández, E. Ley, and M. F. J. Steel, "Benchmark priors for Bayesian model averaging," *J. Econometrics*, vol. 100, no. 2, pp. 381–427, Feb. 2001.

[64] W. Cui and G. E. I., "Empirical Bayes vs. fully Bayes variable selection," *J. Stat. Planning and Inference*, vol. 138, no. 4, pp. 888–900, Apr. 2008.

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

[66] A. Zellner, "Comments on 'Mixtures of g-priors for Bayesian Variable Selection' by F. Liang, R. Paulo, G. Molina, M.A. Clyde and J.O. Berger," Jul. 2008.

[67] 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.

[68] 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.

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

[70] 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.

[71] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "Default Bayesian estimation of the fundamental frequency," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 598–610, Mar. 2013.

[72] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 281–295, Feb. 1996.

[73] M. G. Christensen and J. K. Nielsen, "Joint direction-of-arrival and order estimation in compressed sensing using angles between subspaces," in *Proc. IEEE Workshop on Stat. Signal Process.*, Jun. 2011, pp. 449–452.

[74] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall PTR, Mar. 1993.

Jesper Kjær Nielsen (S'12–M'13) was born in Struer, Denmark, in 1982. He received the B.Sc., M.Sc. (Cum Laude), and Ph.D. degrees in electrical engineering from Aalborg University, Aalborg, Denmark, in 2007, 2009, and 2012, respectively.

He is currently with the Department of Electronic Systems, Aalborg University, and Bang & Olufsen as an industrial postdoctoral researcher. He has been a Visiting Scholar at the Signal Processing and Communications Laboratory, University of Cambridge and at the Department of Computer Science, University of Illinois at Urbana-Champaign. His research interests include spectral estimation, (sinusoidal) parameter estimation as well as statistical and Bayesian methods for signal processing.

Mads Græsbøll Christensen (S'00–M'05–SM'11)) was born in Copenhagen, Denmark, in March 1977. He received the M.Sc. and Ph.D. degrees in 2002 and 2005, respectively, from Aalborg University (AAU) in Denmark, where he is also currently employed at the Dept. of Architecture, Design & Media Technology as Professor in Audio Processing. At AAU, he is head of the Audio Analysis Lab which conducts research in audio signal processing.

He was formerly with the Dept. of Electronic Systems, Aalborg University and has been a Visiting Researcher at Philips Research Labs, ENST, UCSB, and Columbia University. He has published more than 100 papers in peer-reviewed conference proceedings and journals as well as 2 research monographs. His research interests include digital signal processing theory and methods with application to speech and audio, in particular parametric analysis, modeling, enhancement, separation, and coding.

Prof. Christensen has received several awards, including an ICASSP Student Paper Award, the Spar Nord Foundation's Research Prize for his Ph.D. thesis, a Danish Independent Research Council Young Researcher's Award, and the Statoil Prize, as well as grants from the Danish Independent Research Council and the Villum Foundation's Young Investigator Programme. He is an Associate Editor for IEEE Transactions on Audio, Speech, and Language Processing and has previously served as an Associate Editor for IEEE Signal Processing Letters.

A. Taylan Cemgil (M'04) received his Ph.D. (2004) from Radboud University Nijmegen, the Netherlands. Between 2004 and 2008 he worked as a postdoctoral researcher at Amsterdam University and the Signal Processing and Communications Lab., University of Cambridge, UK. He is currently an associate professor of Computer Engineering at Boğaziçi University, Istanbul, Turkey. He is a member of the IEEE Machine Learning for Signal Processing Technical Committee and an associate editor of IEEE Signal processing Letters and Digital Signal Processing. His research interests are in Bayesian statistical methods, approximate inference, machine learning and audio signal processing.

Søren Holdt Jensen (S'87–M'88–SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995. Before joining the Department of Electronic Systems of Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd, Copenhagen, Denmark; the Electronics Institute of the Technical University of Denmark; the Scientific Computing Group of Danish Computing Center for Research and Education (UNI•C), Lyngby; the Electrical Engineering Department of Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation (CPK) of Aalborg University.

He is Full Professor and is currently heading a research team working in the area of numerical algorithms, optimization, and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications.

Prof. Jensen was an Associate Editor for the IEEE Transactions on Signal Processing, Elsevier Signal Processing and EURASIP Journal on Advances in Signal Processing, and is currently Associate Editor for the IEEE Transactions on Audio, Speech and Language Processing. He is a recipient of an European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section and the IEEE Denmark Section's Signal Processing Chapter. He is member of the Danish Academy of Technical Sciences and was in January 2011 appointed as member of the Danish Council for Independent Research—Technology and Production Sciences by the Danish Minister for Science, Technology and Innovation.