Estimation of sinusoidal regression models by stochastic complexity*

Ciprian Doru Giurcăneanu
Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FIN-33101 Tampere, Finland
E-mail:ciprian.giurcaneanu@tut.fi

Abstract

Stochastic complexity (SC), or equivalently, the negative logarithm of the NML (Normalized Maximum Likelihood) was proven to be successful for the estimation of model structure in the linear quadratic regression problem. Recently, the results have been extended to autoregressive (AR) and autoregressive moving average (ARMA) models, whereas most of the information theoretic methods currently applied for determining the number of sine-waves in additive Gaussian noise still rely on asymptotic two-terms formulae where the first term is given by the minus maximum log-likelihood, and the second one is a penalty coefficient that depends on the number of parameters and the sample size. Additionally, the noise is assumed to be white, which is not realistic in most of the practical applications. Our main purpose is to apply sharper approximations of SC for estimating the number of sinusoidal terms in a time series contaminated by AR noise. This is known to be challenging because we have to solve a mixed-spectrum estimation problem. We elaborate on two different SC criteria that involve the Fisher information matrix (FIM) of the investigated model. For small and moderate sample sizes, the experimental results show that SC compares favorably with other well-known criteria such as: Bayesian information criterion (BIC), corrected Kullback information criterion (KICc) and the generalized Akaike information criterion (GAIC).

1 Introduction and preliminaries

We address the estimation of the number of sinusoids observed in additive noise with unknown correlation structure. To formulate the problem, we consider the data model

$$\begin{aligned} y_t &= x_t + e_t,\ t \in \{0, \dots, N-1\}, \\ x_t &= \sum_{k=1}^{K} \alpha_k \cos(\omega_k t + \phi_k), \end{aligned} \tag{1}$$

*The contribution extends the results of the paper "Stochastic complexity for the estimation of sine-waves in colored noise", authored by C.D. Giurcaneanu and presented at ICASSP 2007, Honolulu, Hawaii, USA.

where y_t denotes the measurements, x_t is the noise-free signal and e_t is the colored Gaussian noise.

To ensure the identifiability of the parameters, we assume as usual that the amplitudes α_k are strictly positive and the frequencies ω_k belong to the interval $(0, \pi)$ [1]. The frequencies are distinct and, without loss of generality, $\omega_1 < \cdots < \omega_K$. Both the amplitudes and the frequencies are non-random parameters that will be estimated from the available measurements.

Two different hypotheses will be considered for modeling the phases $\phi_k \in [-\pi, \pi)$: H_{dp} - the phases are unknown deterministic constants; H_{rp} - the phases are independent and uniformly distributed random variables that are also independent of e_t. For both assumptions, the statistical properties of y_t have been investigated in previous studies, and more details can be found, for example, in [2].

In line with the approach from [2][3][4] and the references therein, we model the noise e_t as a stable autoregressive (AR) process with order M:

$$e_t = -\sum_{m=1}^{M} a_m e_{t-m} + w_t, \tag{2}$$

where w_t is a sequence of independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and variance τ. Since we consider only the case of real sinusoids in real AR noise, we emphasize that the white random process w_t and the coefficients a_m, $1 \le m \le M$, are real-valued.

When the noise is white, or equivalently $M = 0$, it is well-known the definition of the local SNR for the k-th sinusoid: $\mathrm{SNR}_k = \dfrac{\alpha_k^2/2}{\tau}$. An extension of this definition, namely $\mathrm{SNR}_k = \dfrac{\alpha_k^2/2}{|H(\exp(j\omega_k))|^2}$, was also introduced in the literature [5] for the case when the additive noise is modeled as the output of an exponentially stable and invertible linear filter $H(q^{-1})$ whose input is a sequence of i.i.d. Gaussian random variables with zero mean and variance τ. We note that q^{-1} is the unit delay operator and $j = \sqrt{-1}$. For the AR noise defined in equation (2), we get immediately

$$\mathrm{SNR}_k = \frac{\alpha_k^2/2}{\tau/|A(\omega_k)|^2}, \tag{3}$$

where $A(\omega_k) = 1 + \sum_{m=1}^{M} a_m \exp(-jm\omega_k)$. A similar formula can be written without difficulties for the case when the additive noise is a moving average process.

Based on (1) and (2), we observe that the parameters of the model are $\boldsymbol{\theta} = [\boldsymbol{\xi}\ \mathfrak{a}\ \tau]^\top$, where $\boldsymbol{\xi} = [\boldsymbol{\xi}_1^\top \cdots \boldsymbol{\xi}_K^\top]^\top$ with the convention $\boldsymbol{\xi}_k = [\alpha_k\ \omega_k\ \phi_k]^\top$ for the k-th sine-wave. The notation $\mathfrak{a}$ is employed for the vector of the AR coefficients $[a_1 \cdots a_M]^\top$.

Because the model structure $\gamma = (K, M)$ is not known a priori, we resort to the traditional model selection procedure that comprises two steps:

(a) for all pairs of integers $\gamma = (K, M)$ that satisfy $0 \le K \le K_{max}$ and $0 \le M \le M_{max}$, estimate the model parameters $\hat{\boldsymbol{\theta}}_\gamma$ from the observations $y^N = y_0, \ldots, y_{N-1}$;

(b) evaluate an information theoretic criterion for all γ's considered at the first step, and choose the model structure $\hat{\gamma}$ that minimizes the criterion.

The most popular rules for model selection can be reduced to a common form with two terms: the first one is the minus maximum log-likelihood, and the second one is a penalty coefficient that depends on the number of parameters of the model and, for some criteria, also on the sample size N [6]. In general, the criteria used in practical applications are derived for $N \to \infty$, and the asymptotic approximations could potentially yield false conclusions when the sample size is small or moderate.

During recent years, the advances in stochastic complexity (SC) have led to new exact formulae or to sharper approximations for large classes of models [7][8][9], but the use of the new results in signal processing is scarce. We illustrate next how SC can be applied to estimate the structure for the model of sine-waves in Gaussian AR noise.

The rest of the paper is organized as follows. In Section 2, two different approximative formulae for SC are revisited: one proposed by Rissanen in [8], and another one introduced by Qian and Künsch in [7]. As the computation of the Rissanen sharp approximation is difficult for the sinusoidal regression model, we focus on the Qian and Künsch formula, and we investigate its properties in Section 3. Because the SC expression involves the determinant of the Fisher information matrix (FIM), the calculation of FIM is addressed in Section 4. SC and three other well-known model selection criteria are compared in Section 5 to evaluate their performances in estimating the number of sinusoids from simulated data.

2 SC for sine-waves in AR noise

We focus on the expression of SC for the class of the density functions $\{f(y^N;\boldsymbol{\theta})\}$ defined by the equations (1) and (2). For the maximum likelihood (ML) estimates we employ the notation $\hat{\boldsymbol{\theta}}(y^N)$ whenever it is necessary to emphasize on the data set. If it is clear from the context which measurements are used for estimation, then the simpler notation $\hat{\boldsymbol{\theta}}$ is preferred to $\hat{\boldsymbol{\theta}}(y^N)$. Therefore $\ln f(y^N;\hat{\boldsymbol{\theta}}) = \ln f(y^N;\hat{\boldsymbol{\theta}}(y^N))$ is the maximum log-likelihood, Θ denotes the parameter space, and $\mathbf{J}_N(\boldsymbol{\theta}) = E\left[-\dfrac{\partial^2 \ln f(y^N;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]$ is the Fisher information matrix (FIM). The Normalized Maximum Likelihood (NML) density function is given by [10][8],

$$\hat{f}(y^N;K,M) = \frac{f(y^N;\hat{\boldsymbol{\theta}}(y^N))}{\int_{x^N:\hat{\boldsymbol{\theta}}(x^N)\in\Theta} f(x^N;\hat{\boldsymbol{\theta}}(x^N))\mathrm{d}x^N},$$

and the stochastic complexity is defined as

$$\mathrm{SC}(y^N;K,M) = \ln\left(1/\hat{f}(y^N;K,M)\right).$$

The NML criterion has two important optimality properties [11] that recommend it to be used as a yardstick in model selection. The application of the NML criterion is appealing, but its computation is not very easy for all classes of models. Under mild assumptions on ML estimates, SC is approximated in [8] with a formula that involves the integral of the squared root of the

FIM determinant. The approximation is valid only if FIM divided by N, the number of samples, has a finite limit as $N \to \infty$. The condition is verified for most of the models used in signal processing, but not for the sinusoidal regression model [6]. We show next how the results from [8] can be extended to the sinusoidal regression model, and we also point out the difficulties with evaluating the integral term. Due to the troubles with the integral, we resort to another SC approximation that was introduced by Qian and Künsch in [7].

2.1 Sharp approximations of SC

As the Rissanen formula involves the asymptotic FIM, the following result is very useful for our application: when N is large, under both $\mathrm{H_{dp}}$ and $\mathrm{H_{rp}}$, $\mathbf{J}_N(\boldsymbol{\theta})$ is block-diagonal such that the block $\mathbf{J}_N(\boldsymbol{\xi}_k)$ corresponds to the parameters of the k-th sine-wave and the block $\mathbf{J}_N(\mathfrak{a}, \tau)$ corresponds to the parameters of the AR noise [2][5]. More precisely, we have

$$\mathbf{J}_N(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{J}_N(\boldsymbol{\xi}_1) & & & \\ & \ddots & & \\ & & \mathbf{J}_N(\boldsymbol{\xi}_K) & \\ & & & \mathbf{J}_N(\mathfrak{a}, \tau) \end{bmatrix} \tag{4}$$

where

$$\mathbf{J}_N(\boldsymbol{\xi}_k) = \mathbf{Q}_N \mathbf{G}(\boldsymbol{\xi}_k, \mathfrak{a}, \tau) \mathbf{Q}_N, \tag{5}$$

$$\mathbf{Q}_N = \begin{bmatrix} N^{1/2} & 0 & 0 \\ 0 & N^{3/2} & 0 \\ 0 & 0 & N^{1/2} \end{bmatrix} \quad \text{and} \quad \mathbf{G}(\boldsymbol{\xi}_k, \mathfrak{a}, \tau) = \mathrm{SNR}_k \begin{bmatrix} 1/\alpha_k^2 & 0 & 0 \\ 0 & 1/3 & 1/2 \\ 0 & 1/2 & 1 \end{bmatrix}. \tag{6}$$

Here SNR_k denotes the local SNR for the k-th sinusoidal component and its formula is given in (3). The entries of $\mathbf{J}_N(\mathfrak{a}, \tau)$ are not influenced by the parameters $\boldsymbol{\xi}$, hence $\mathbf{J}_N(\mathfrak{a}, \tau)$ is the same as in the pure AR case. Based on results from [12], we can write

$$\mathbf{J}_N(\mathfrak{a}, \tau) = \frac{N}{\tau} \begin{bmatrix} \mathbf{R}(\mathfrak{a}) & \mathbf{0} \\ \mathbf{0} & 1/(2\tau) \end{bmatrix}, \tag{7}$$

where $\mathbf{R}(\mathfrak{a}) = \begin{bmatrix} r_0 & \cdots & r_{M-1} \\ \vdots & \ddots & \vdots \\ r_{M-1} & \cdots & r_0 \end{bmatrix}$ is the covariance matrix of the AR process defined in (2).

Additionally we define the diagonal matrix $\mathbf{C}_N = \begin{bmatrix} \mathbf{I}_K \otimes \mathbf{Q}_N & \mathbf{0} \\ \mathbf{0} & N^{1/2} \times \mathbf{I}_{M+1} \end{bmatrix}$, where the symbol $\otimes$ denotes the Kronecker product, and for a strictly positive integer p, $\mathbf{I}_p$ is the $p \times p$ identity matrix. We adopt the convention that $\mathbf{0}$ denotes a null vector/matrix of appropriate dimensions. Based on (4)-(7), we note that $\lim_{N\to\infty} \frac{1}{N}\mathbf{J}_N(\boldsymbol{\theta})$ is not finite, whereas $\mathbf{J}(\boldsymbol{\theta}) = \lim_{N\to\infty} \mathbf{C}_N^{-1}\mathbf{J}_N(\boldsymbol{\theta})\mathbf{C}_N^{-1}$ is finite [6]. Moreover, the ML estimates satisfy the Central Limit Theorem: the distribution of $\mathbf{C}_N(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to the Gaussian distribution of mean zero and covariance $\mathbf{J}(\boldsymbol{\theta})^{-1}$ [4]. These properties allow us to extend the results from [8] to the sinusoidal regression model for which the SC formula is given by

$$\mathrm{SCr}(y^N; K, M) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + \frac{5K+M+1}{2} \ln \frac{N}{2\pi} + \ln \int_{\Theta} |\mathbf{J}(\boldsymbol{\theta})|^{1/2} \mathrm{d}\boldsymbol{\theta} + o(1). \tag{8}$$

We use the notation SCr to differentiate this particular approximation of SC by other formulae that will be discussed later.

Remark that $\mathbf{J}(\boldsymbol{\theta})$ is a block-diagonal matrix, and $\frac{1}{\tau}\mathbf{R}(\mathfrak{a}, \tau)$ is the block corresponding to the parameters of the Gaussian autoregressive noise. Computing the integral of $\left|\frac{1}{\tau}\mathbf{R}(\mathfrak{a}, \tau)\right|^{1/2}$ over the parameter space is a problem that arose also in the context of order estimation for AR processes [13]. Since it is hard to find a closed-form expression of the integral, the authors of [13] resorted to Monte Carlo techniques for its evaluation. Our task here is even more difficult because the other blocks of $\mathbf{J}(\boldsymbol{\theta})$ must be also considered when calculating the integral term. Hence the computational burden discourages us to apply formula (8) for estimating the number of sine-waves in Gaussian autoregressive noise. We show next that SC expression (8) becomes simpler when the noise is white ($M = 0$). In this case, $\boldsymbol{\theta} = [\boldsymbol{\xi}^\top \; \tau]^\top$, and elementary calculations lead to

$$\mathrm{SCr}(y^N; K, 0) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + \frac{5K+1}{2} \ln \frac{N}{2\pi} + \ln \int_{\Theta} \frac{1}{2^{1/2} 96^{K/2}} \frac{\prod_{k=1}^{K} \alpha_k^2}{\tau^{(3K+2)/2}} \mathrm{d}\boldsymbol{\theta} + o(1). \tag{9}$$

To ensure that the integral term is finite, we have to assume that all amplitudes have an upper bound, $\alpha_k < \alpha_{max} < \infty$, and the noise variance has a strictly positive lower bound, $\tau > \tau_{min} > 0$. Once these conventions are adopted, the estimated number of sine-waves will depend on α_{max} and τ_{min}. Note that α_{max} and τ_{min} are just arbitrary values if we do not have a priori knowledge on the analyzed signals. The troubles with the computation of SCr make us to prefer the SC formula that was derived in [7]:

$$\mathrm{SC}(y^N; K, M) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + \ln |\mathbf{J}_N(\hat{\boldsymbol{\theta}})|^{1/2} + \sum_{i=1}^{3K+M+1} \ln(|\hat{\theta}_i| + N^{-1/4}) \tag{10}$$

A similar approximation of SC was already utilized in [14] to estimate K, the number of sinusoids. In [14], the noise variance τ is treated as a nuisance parameter in the sense that the code length to describe it is not included in the SC formula. Here we consider in (10) the cost for transmitting the value of the τ parameter, and this is the main difference between our approach and the one from [14]. We check in the Appendix how the conditions for the derivations from [7] are fulfilled for the model of sinusoids in Gaussian AR noise. We also give in the Appendix more details on the accuracy of the approximation in formula (10).

In the next Section, we investigate the asymptotic behavior of the SC criterion and we show its relation with well-known selection rules like Bayesian information criterion (BIC) [15], Minimum Description Length (MDL) [16], and the *maximum a posteriori* (MAP) probability criterion [17]. During the asymptotic analysis, we check also the necessary conditions for the consistency [18] of SC. For small and moderate sample sizes, we draw a parallel between SC and two other recently introduced model selection methods: Conditional Model Estimator (CME) [19] and the Exponentially Embedded Families (EEF) [20].

3 Some properties of SC and its relation to other model selection criteria

3.1 BIC, MDL and MAP

Based on the results from the Appendix, we obtain readily the well-known asymptotic identity $\lim_{N\to\infty} \ln|\mathbf{J}_N(\hat{\boldsymbol{\theta}})|^{1/2} = \frac{5K+M+1}{2}\ln N$, and it is easy to notice that the sum of the first two terms in SC is equivalent with the Bayesian information criterion:

$$\mathrm{BIC}(y^N; K, M) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + \frac{5K+M+1}{2}\ln N. \tag{11}$$

More details on the derivation of BIC can be found in [6] and the references therein. In [21], it is investigated the possibility of improving the performances of BIC for small and moderate sample sizes by considering two terms that are neglected in the asymptotic formula (11): the first one involves the logarithm of the determinant of the observed FIM, and the second one is mainly determined by a prior over the family of the analyzed models. As the sinusoidal regression model is not discussed in [21], we restrict our interest to the celebrated BIC selection rule (11), and we do not consider in simulations any sharp approximation of the Bayesian information criterion. We mention for completeness that formula (11) was also obtained in [3] as a crude version of the MDL, and its consistency was demonstrated in the same study. In [17], the use of the MAP methodology in conjunction with asymptotic approximations led also to (11) for the particular case of white noise.

3.2 A short note on the consistency of SC for $M = 0$

For ease of presentation we investigate the consistency of the criterion $\mathrm{SC}'(y^N; K, 0) = \mathrm{SC}(y^N; K, 0) - \frac{1}{2}\ln N$. It is evident that SC′ and SC are equivalent selection rules because $\frac{1}{2}\ln N$ is independent of K. We focus on the last term in (10), and for simplicity we assume $M = 0$. If zero does not belong to the domain of the parameter θ_i, then $\hat{\theta}_i \neq 0$ and $\ln(|\hat{\theta}_i| + N^{-1/4})$ is much smaller than $\frac{1}{2}\ln N$ [7]. Hence the term $\ln(|\hat{\theta}_i| + N^{-1/4})$ becomes important only when $\hat{\theta}_i \approx 0$. Since among the $\boldsymbol{\xi}$ parameters only the phases can be equal to zero, the penalty term in SC′ formula takes values between $\frac{9K}{4}\ln N$ and $\frac{5K}{2}\ln N$ when N is large. Based on formula (10), we can write

$$\mathrm{SC}'(y^N; K, 0) = -\ln f(y^N; \hat{\boldsymbol{\theta}}) + K\zeta(N, \hat{\boldsymbol{\theta}}),$$

and asymptotically $\frac{9}{4}\ln N \leq \zeta(N, \hat{\boldsymbol{\theta}}) \leq \frac{5}{2}\ln N$. Thus $\lim_{N\to\infty} \frac{\zeta(N,\hat{\boldsymbol{\theta}})}{N} = 0$ and $\liminf_{N\to\infty} \frac{\zeta(N,\hat{\boldsymbol{\theta}})}{\ln N} > 1$. If supplementarily the model (1) verifies $\frac{\omega_k}{2\pi} \in \left\{\frac{1}{N}, \cdots, \frac{(N-1)/2}{N}\right\} \forall k \in \{1, \ldots, K\}$, all the conditions for the application of the Theorem from [18] are satisfied. We select $\hat{K}$ to be the minimum nonnegative integer for which $\mathrm{SC}'(y^N; K, 0) < \mathrm{SC}'(y^N; K+1, 0)$, and the Theorem guarantees that $\hat{K}$ converges almost surely to the true number of sinusoids.

3.3 CME, EEF and an example from [20]

In [20], it was shown that using the determinant of FIM as a penalty term could lead to modest results when the sample size is small. As the example from [20] involves sinusoidal signals, we briefly discuss it in the sequel: the noise-free signal x_t is generated like in (1) by a sum of $K = 3$ sine-waves whose parameters are $\boldsymbol{\xi}_1 = [1\ 0.2\pi\ 0]^\top$, $\boldsymbol{\xi}_2 = [1\ 0.22\pi\ 0]^\top$ and $\boldsymbol{\xi}_3 = [1\ 0.24\pi\ 0]^\top$. The white noise e_t is Gaussian with variance $\tau = 10$, and the selection is restricted to the class of nested models $\mathcal{M}_\kappa$, $\kappa \in \{2, 4, \ldots, 16\}$, defined by

$$\mathcal{M}_\kappa \ :\ y_t = \sum_{k=1}^{\kappa/2} \alpha_k \cos(\omega_k t + \phi_k) + e_t,\ \ t \in \{0, \ldots, N-1\},$$

where $\omega_k = 2\pi\left(0.1 + \frac{k-1}{100}\right)$. Since the frequencies are known, $\mathcal{M}_\kappa$ reduces to the linear regression for which the observation matrix has the expression

$$\mathbf{H}_\kappa = \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \cos(\omega_1(N-1)) & \sin(\omega_1(N-1)) & \cdots & \cos(\omega_{\kappa/2}(N-1)) & \sin(\omega_{\kappa/2}(N-1)) \end{bmatrix},$$

and the vector of the unknown parameters is given by $\mathbf{v}_\kappa = [A_1\ B_1 \cdots A_{\kappa/2}\ B_{\kappa/2}]^\top$, where $A_k = \alpha_k \cos\phi_k$ and $B_k = -\alpha_k \sin\phi_k$ for all $k \in \{1, \ldots, \kappa/2\}$. The noise variance τ is assumed to be known. Remark that the number of parameters for the $\mathcal{M}_\kappa$ model is κ. Applying the CME criterion is equivalent with choosing the model $\mathcal{M}_{\hat{\kappa}}$ that minimizes [19]

$$\mathrm{CME}(y^N; \kappa) = \frac{\mathrm{RSS}_\kappa}{\tau} + \ln\left|\frac{\mathbf{H}_\kappa^\top \mathbf{H}_\kappa}{2\pi\tau}\right|,$$

where RSS_κ is the residual sum of squares obtained when fitting the $\mathcal{M}_\kappa$ model to the observations y^N. In [20], it was utilized the approximation $\mathbf{H}_\kappa^\top \mathbf{H}_\kappa \approx (N/2)\mathbf{I}_\kappa$ to show that the second term in the equation above is negative when $N < 125$, thus the penalty term of the CME criterion decreases when κ increases. Since for the models considered in this example, $\frac{1}{\tau}\mathbf{H}_\kappa^\top \mathbf{H}_\kappa$ coincides with the FIM [1], Kay concluded in [20] that all criteria whose penalty factor is given by the determinant of FIM will always choose the most complex model when the sample size is small or moderate. To circumvent such difficulties, he introduced the EEF criterion that, for linear regression models, amounts to select the $\mathcal{M}_{\hat{\kappa}}$ model that minimizes

$$\mathrm{EEF}(y^N; \kappa) = \left[-Q_\kappa + \kappa\left(\ln\frac{Q_\kappa}{\kappa} + 1\right)\right] u\left(\frac{Q_\kappa}{\kappa} - 1\right), \tag{12}$$

where $Q_\kappa = \dfrac{\|\mathbf{H}_\kappa \hat{\mathbf{v}}_\kappa\|^2}{\tau}$, the entries of $\hat{\mathbf{v}}_\kappa$ are the ML estimates of the parameters, and $u(\cdot)$ is the step unit function [20].

We use the same example to investigate if similar drawbacks appear when the model selection relies on the SC criterion. The FIM-based SC approximation [8] was computed in [22] for the linear regression case, and it involves the ranges of the parameters, which is not convenient as we have already pointed out in Section 2.1. Fortunately we do not need to resort to such an approximation because Rissanen gave in [9] a very elegant solution to the problem of evaluating SC for the linear regression model. For the analyzed example, we prefer to apply the result from [9] in the form that was worked out in [23]:

$$\mathrm{SClr}(y^N;\kappa) = \frac{1}{\tau}\sum_{i=0}^{N-1} y_i^2 - Q_\kappa + \kappa\left(\ln\frac{Q_\kappa}{\kappa}+1\right) + \ln\kappa,$$

where the notations are the same like in (12). We observe that unlike the CME criterion, SClr does not contain the term given by the determinant of FIM. Moreover, the expressions of SClr and EEF are very similar. It is easy to note that $-Q_\kappa$ decreases with κ. Let us consider first the case when $\frac{Q_\kappa}{\kappa} > 1$. In general, the term $\kappa\left(\ln\frac{Q_\kappa}{\kappa}+1\right)$ increases with κ [20], hence it is a penalty term for both EEF and SClr. Remark that in this case, due to the $\ln\kappa$ term, the penalty will be more stringent for SClr than for EEF. Formula (12) can be re-written as $\mathrm{EEF}(y^N;\kappa) = \kappa h(Q_\kappa/\kappa)$, where $h(x) = -x + \ln x + 1, \forall x \in (0,\infty)$. Because $h(x)$ is strictly negative for $x > 1$, the criterion EEF has the same property. Whenever $\frac{Q_\kappa}{\kappa} \leq 1$, $\mathrm{EEF}(y^N;\kappa)$ takes value zero, and consequently the model $\mathcal{M}_\kappa$ will not be selected. For SClr, if $\frac{Q_\kappa}{\kappa}$ is small, the term $\kappa\left(\ln\frac{Q_\kappa}{\kappa}+1\right)$ could become negative and $\ln\kappa$ will remain the only penalty term.

For Gaussian linear regression with known noise variance, another SC criterion was derived in [24] by using the universal mixture model instead of the NML:

$$\mathrm{SChy}(y^N;\kappa) = \frac{1}{2\tau}\sum_{i=0}^{N-1} y_i^2 + \frac{1}{2}\left[-Q_\kappa + \kappa\left(\ln\frac{Q_\kappa}{\kappa}+1\right) + \ln N\right] u\left(\frac{Q_\kappa}{\kappa}-1\right). \tag{13}$$

As it was pointed out in [24], SChy coincides up to the $\frac{1}{2}\ln N$ additive term with the empirical Bayesian selection rule proposed in [25]. Comparing (12) and (13) we also note that EEF and SChy are essentially the same.

4 Computational issues

The use of (10) is very appealing from computational viewpoint, but it was already pointed out in [7] that (10) is not invariant under re-parametrization. Due to this reason, we prefer to use as parameters for the AR noise the magnitudes and the angles of the poles instead of the coefficients.

More precisely, let us assume that the poles of the AR noise model are $g_1,\ldots,g_M$: if the poles $g_1,\ldots,g_{M_1}$ are real-valued, then the pure complex poles $g_{M_1+1},\ldots,g_M$ occur in complex conjugate pairs because the coefficients $\mathfrak{a}$ are real-valued. Instead of $\boldsymbol{\theta} = [\boldsymbol{\xi}\ \mathfrak{a}\ \tau]^\top$, we will apply the parametrization $\boldsymbol{\eta} = [\boldsymbol{\xi}\ \mathfrak{g}\ \tau]^\top$, where $\mathfrak{g} = [g_1 \ldots g_{M_1}\ |g_{M_1+1}|\ \psi_{g_{M_1+1}} \cdots |g_{M-1}|\ \psi_{g_{M-1}}]^\top$, and for a complex pole g_i, the symbol ψ_{g_i} denotes its angle. Remark the range of the entries of $\mathfrak{g}$: we have $g_i \in (-1,1)$ for $1 \leq i \leq M_1$, and for the rest of the parameters $|g_i| \in (0,1)$ and $\psi_{g_i} \in (0,\pi)$.

SC	Hypothesis	$\mathbf{J}_N(\boldsymbol{\xi})$	$\mathbf{J}_N(\mathfrak{g})$
SCp	$\mathrm{H_{rp}}$	exact	exact
SCa	$\mathrm{H_{rp}}/\mathrm{H_{dp}}$	asymptotic	asymptotic
SCe	$\mathrm{H_{dp}}$	exact	exact

Table 1: Nomenclature for SC when various formulae for FIM are used in calculations.

To calculate the determinant of the FIM with the new parametrization, we use the general result on the transformation of parameters [1] in conjunction with the result of equation (15) from [2]. For writing the equations in a more compact form, we define the $(M+1)\times(M+1)$ matrix $\left[\dfrac{D(\mathfrak{a},\tau)}{D(\mathfrak{g},\tau)}\right]$ whose (m,n)-th element is $\dfrac{\partial \mathfrak{a}_m}{\partial \mathfrak{g}_n}$ if $1 \leq m,n \leq M$, it is one if $m=n=M+1$, and otherwise takes value zero. Next we obtain the following identities:

$$\begin{aligned} |\mathbf{J}_N(\boldsymbol{\eta})| &= \begin{vmatrix} \mathbf{I}_{3K} & \mathbf{0} \\ \mathbf{0} & \left[\frac{D(\mathfrak{a},\tau)}{D(\mathfrak{g},\tau)}\right] \end{vmatrix}^{\top} \begin{vmatrix} \mathbf{J}_N(\boldsymbol{\xi}) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_N(\mathfrak{a},\tau) \end{vmatrix} \begin{vmatrix} \mathbf{I}_{3K} & \mathbf{0} \\ \mathbf{0} & \left[\frac{D(\mathfrak{a},\tau)}{D(\mathfrak{g},\tau)}\right] \end{vmatrix} \\ &= |\mathbf{J}_N(\boldsymbol{\xi})||\mathbf{J}_N(\mathfrak{g},\tau)| \end{aligned}$$

The block $\mathbf{J}_N(\boldsymbol{\xi})$ that corresponds to the signal parameters can be evaluated with the fast algorithms from [2]: the exact $\mathbf{J}_N(\boldsymbol{\xi})$ is different for $\mathrm{H_{dp}}$ and $\mathrm{H_{rp}}$, but the asymptotic $\mathbf{J}_N(\boldsymbol{\xi})$ has the same form under both hypotheses. This asymptotic form is well-known [5], and it is also given in Section 2.1. $\mathbf{J}_N(\mathfrak{a},\tau)$ has the same expression as in the pure AR case, and for its calculation we resort to the exact and the asymptotic formulae from [12]. The conversion from $\mathbf{J}_N(\mathfrak{a},\tau)$ to $\mathbf{J}_N(\mathfrak{g},\tau)$ can be easily performed with the results from [26]. A discussion on the asymptotic form of $\mathbf{J}_N(\mathfrak{g},\tau)$ can be found in [13].

Applying the exact or asymptotic formulae for $\mathbf{J}_N(\boldsymbol{\xi})$ and $\mathbf{J}_N(\mathfrak{g},\tau)$ leads to various expressions for SC. In Table 1, we explain the nomenclature for SC when FIM in (10) is evaluated with various formulae.

For better understanding the differences between SCp, SCa and SCe we resort to one of the examples used in [2] to analyze the Cramer-Rao bound (CRB). Let us consider the case of one single sinusoid ($K=1$) in AR noise with order $M=2$. We choose $\alpha_1=1$, $\omega_1=\pi/2$, the modulus of the AR poles is $|g_1|=0.9$, and the sample size is $N=35$. The angle ψ_{g_1} takes values between 0.02 and $(\pi-0.02)$, and the variance τ is selected such that to keep constant $\mathrm{SNR}_1 =$ 3 dB. Evaluating the differences between SCp, SCa and SCe reduces to calculate $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ with various formulae. Because under $\mathrm{H_{dp}}$ the exact $\mathbf{J}_N(\boldsymbol{\xi})$ depends on the phase ϕ_1, for each ψ_{g_1} we compute $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ for sixty different values of ϕ_1 that are equally spaced in $[-\pi,\pi)$, and the largest ($\bigtriangledown$) and the smallest ($\bigtriangleup$) results are plotted in Figure 1. We plot in the same Figure the values of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ used in the calculation of SCp (dash-dot line) and SCa (continuous line). For sake of comparison, we draw also a horizontal line that corresponds to $\frac{5K+M+1}{2}\ln N$. We can easily extend the conclusions on CRB drawn in [2], by observing the significant difference between the asymptotic approximation of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ and its exact value when the line spectrum is close to the spectral peak of the noise. Remark also in Figure 1 that

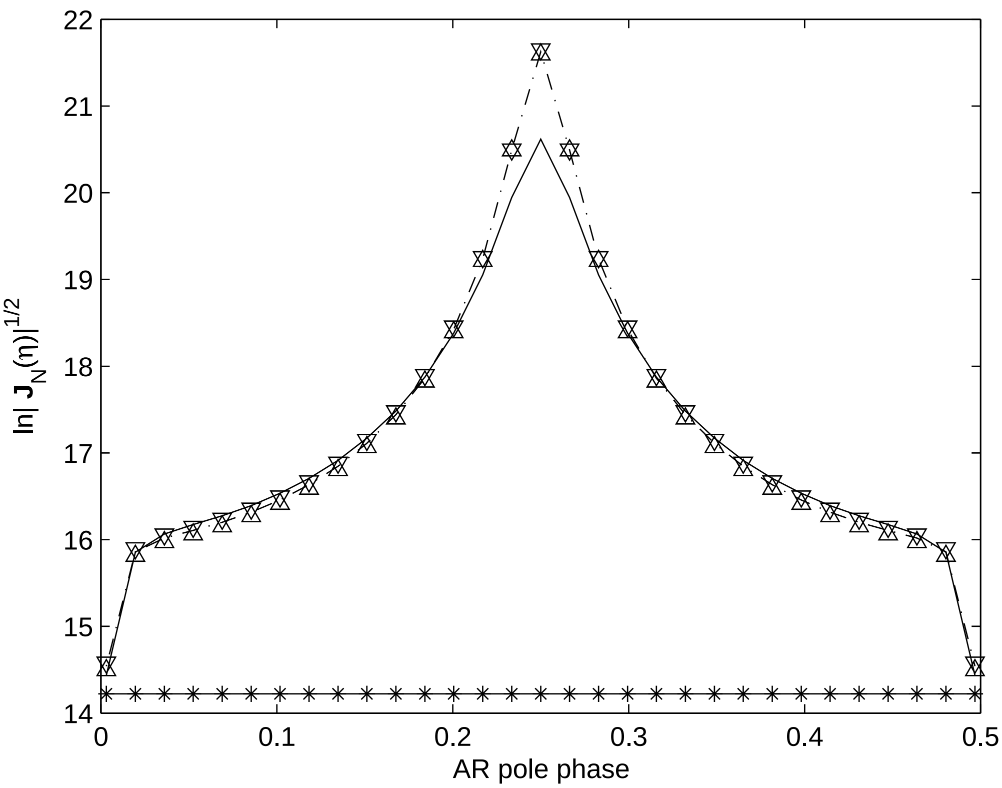

Figure 1: The term $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ versus the phase ψ_{g_1} of the AR pole when the sample size is $N = 35$. In the case of the SCe formula, $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ is calculated for sixty different values of $\phi_1 \in [-\pi, \pi)$, and the largest ($\bigtriangledown$) and the smallest ($\bigtriangleup$) results are plotted. The dash-dot line and the continuous line are for the values of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ as they are used in the evaluation of the SCp and SCa, respectively. The horizontal line with a $\star$ at each data point corresponds to $\frac{5K+M+1}{2}\ln N$.

the value of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ used to compute SCp is approximately equal with the average of the maximum and the minimum of $\ln|\mathbf{J}_N(\boldsymbol{\eta})|^{1/2}$ employed in the calculation of SCe.

In the next Section we investigate how the structure estimation performances of SC are influenced by the use of various formulae for FIM.

5 Experimental results

In all the examples presented next, we resort to the RELAX algorithm that performs a decoupled parameter estimation for the sinusoids and the AR noise [4]. In our simulations we have used for the implementation of RELAX the Matlab functions that are publicly available at *http://www.uni-kassel.de/fb16/hfk/neu/toolbox*.

Asymptotically both RELAX and the maximum likelihood (ML) yield statistically efficient estimates, and the use of RELAX is recommended due to its lower computational burden [4][5].

For $\gamma = (K, M)$, let $\hat{\boldsymbol{\xi}}_k$ be the parameters of the k-th sinusoid estimated with RELAX. We denote $\hat{e}_t = y_t - \sum_{k=1}^{K} \hat{\alpha}_k \cos(\hat{\omega}_k t + \hat{\phi}_k)$, and let $\hat{\mathbf{a}}$ be the coefficients of the AR noise determined from the sequence $\hat{e}_0, \ldots, \hat{e}_{N-1}$. We further define the residual sum of squares as $\mathrm{RSS}_\gamma = \sum_{t=0}^{N-1} \left[\hat{e}_t + \sum_{m=1}^{M} \hat{a}_m \hat{e}_{t-m}\right]^2$, with the convention that $\hat{e}_t = 0$ for $t < 0$.

The performances of SC are compared in our simulations with BIC (11) and two other criteria: GAIC and KICc. GAIC is a generalized Akaike Information Criterion that was traditionally used in conjunction with the RELAX algorithm [4]. It seeks for the model structure γ that minimizes

$$\mathrm{GAIC}(y^N; K, M) = N \ln \mathrm{RSS}_\gamma + 8(3K + M + 1)\ln(\ln N).$$

KICc was derived in [27] as a unbiased Kullback Information Criterion for linear regression models with i.i.d. Gaussian noise. Since then its application was extended also to other classes of models, see for example [28] and the references therein. Applying KICc is equivalent with selecting the model structure γ that minimizes [27]

$$\mathrm{KICc}(y^N; K, M) = -2 \ln f(y^N; \hat{\boldsymbol{\theta}}) + 2\frac{(\kappa+1)N}{N-\kappa-2} - N\psi\left(\frac{N-\kappa}{2}\right) + N \ln \frac{N}{2}, \qquad (14)$$

where $\kappa = K + M$ and $\psi(\cdot)$ is the *digamma* function [29]. In SC (10), BIC (11) and KICc (14), $-\ln f(y^N; \hat{\boldsymbol{\theta}})$ is evaluated as $\frac{N}{2} \ln \mathrm{RSS}_\gamma$ after discarding the terms that do not depend on γ.

In our settings, the maximum number of sinusoids is $K_{max} = 8$, and the maximum order of the AR process depends on the number of the available measurements: $M_{max} = \lfloor \ln^2 N \rfloor - 1$. The formula for M_{max} is derived from the condition used in [3] to ensure the consistency of the BIC criterion. Supplementarily, each pair (K, M) must verify the inequality $3K + M < N - 2$ to be a candidate for the model structure.

Examples 1-3 are taken from [3], where the estimation results are reported only for $N \geq 128$. Since our main interest is on small and moderate sample sizes, we evaluate the performances of the information theoretic criteria for $N \in \{25, \ldots, 100\}$ and various levels of the local SNR. In Examples 1-3, we consider $K = 2$ sinusoids whose parameters are $\boldsymbol{\xi}_1 = [2^{1/2}\ 1\ 0]^\top$ and $\boldsymbol{\xi}_2 = [2^{-1/2}\ 2\ 0]^\top$. The additive noise is generated as follows:

Example 1: $e_t = \varepsilon_t$ (white noise),

Example 2: $e_t = -0.81 e_{t-2} + \varepsilon_t$ (autoregressive noise),

Example 3: $e_t = \varepsilon_t + 1.6\varepsilon_{t-1} + 0.64\varepsilon_{t-2}$ (moving average noise),

where ε_t is a sequence of i.i.d. Gaussian random variables with zero mean and variance chosen such that the local SNR's take the desired values.

Example 4 is taken from [4] and modified such that the observations y^N are real-valued. The number of sinusoids is $K = 3$ and their parameters are $\boldsymbol{\xi}_1 = [2\ 0.10\pi\ 0]^\top$, $\boldsymbol{\xi}_2 = [2\ 0.80\pi\ 0]^\top$ and $\boldsymbol{\xi}_3 = [2\ 0.84\pi\ 0]^\top$. The noise is simulated by the autoregressive process $e_t = 0.85 e_{t-1} + \varepsilon_t$, where the significance of ε_t is the same as above.

We focus on the capabilities of the tested criteria to estimate correctly the number of sinusoids K. For the Examples 1-4, we count the number of correct estimates for 100 runs when the local SNR's and the sample size N take various values. The results are reported in Tables 2-5.

SNR$_2$=-3.00 dB								
N	30	40	50	60	70	80	90	100
SCp	41	61	**79**	**86**	**94**	**94**	**97**	**100**
SCa	41	63	76	84	**94**	93	**97**	**100**
SCe	26	52	64	63	78	84	89	89
BIC	25	41	58	68	87	85	90	96
KICc	**54**	**80**	77	74	61	45	44	40
GAIC	3	6	13	22	38	59	65	67
SNR$_2$=-1.00 dB								
N	30	35	40	45	50	60	80	100
SCp	62	72	**88**	**93**	**95**	**99**	98	99
SCa	64	67	71	89	85	91	94	98
SCe	44	59	71	83	82	85	86	96
BIC	43	56	69	78	80	91	**99**	**100**
KICc	**81**	**81**	79	77	78	67	49	38
GAIC	7	19	40	40	53	73	93	**100**
SNR$_2$=0.00 dB								
N	25	30	35	40	45	50	75	100
SCp	60	79	**93**	**96**	**97**	**98**	**98**	99
SCa	60	68	76	76	86	88	94	97
SCe	45	68	75	80	81	90	92	96
BIC	37	51	68	82	86	88	**98**	99
KICc	**75**	**89**	89	92	91	80	56	40
GAIC	6	18	37	52	67	74	97	**100**

Table 2: Example 1: the counts indicate for 100 runs the number of times the number of sinusoids was correctly estimated by each criterion. The best result for each sample size N is represented with bold font.

We note that the estimation results are similar with those reported in [14]. SCp is the best among the SC formulae and its performances are closely followed by SCa. For both SCp and SCa, FIM of the sinusoidal components are decoupled [2], which is a serious computational advantage. From the results reported in [2] together with the outcome of the Example discussed in Section 4, we can draw the conclusion that the shape of the noise spectrum has more influence on SCp than on SCa, and this explains the superiority of the SCp criterion. The performances of SCe are very modest because FIM used in SCe can be ill-conditioned for small and moderate sample size when the number of sinusoids is two or larger [2].

When the sample size N is smaller than 80, SCa is superior to BIC and GAIC. This is a straightforward consequence of the asymptotic approximations applied in the derivations of the BIC and GAIC criteria. KICc estimates for the number of sinusoids are remarkably correct when $N \leq 40$, but the number of correct estimations yield by KICc declines when N increases such that for $N \geq 80$ the reported results are very modest.

SNR$_2$=1.00 dB								
N	30	40	50	60	70	80	90	100
SCp	**26**	**63**	**76**	**81**	**90**	**87**	87	92
SCa	25	53	74	78	87	84	86	92
SCe	24	34	74	79	89	**87**	87	91
BIC	9	41	61	76	89	86	**89**	**93**
KICc	23	36	35	37	42	36	33	43
GAIC	5	12	28	42	64	78	**89**	**93**
SNR$_2$=3.00 dB								
N	30	35	40	45	50	60	80	100
SCp	**50**	**61**	**76**	**96**	**92**	**95**	96	97
SCa	44	52	60	86	83	93	95	97
SCe	41	46	34	50	88	88	96	97
BIC	22	41	53	71	77	91	95	**98**
KICc	29	44	45	38	42	43	46	45
GAIC	7	17	28	38	54	75	**97**	**98**
SNR$_2$=5.00 dB								
N	25	30	35	40	45	50	75	100
SCp	36	**63**	**81**	**88**	**90**	**95**	95	99
SCa	**41**	58	73	51	67	91	93	99
SCe	27	54	57	21	15	89	94	99
BIC	16	28	65	57	77	83	99	**100**
KICc	29	40	57	54	47	52	50	44
GAIC	14	16	32	47	60	72	**100**	**100**

Table 3: Example 2: the performances in estimating the number of sine-waves reported with the same conventions as in Table 2.

We extend our analysis by counting the Type I and Type II errors. Let $f_k = \omega_k/(2\pi)$ and similarly $\hat{f}_k = \hat{\omega}_k/(2\pi)$. Since K and $\hat{K}$ are not necessarily equal, we take $\mathcal{K} = \min(K, \hat{K})$. We select the indices $\{i_1, \ldots, i_{\mathcal{K}}\} \subseteq \{1, \ldots, K\}$ and $\{j_1, \ldots, j_{\mathcal{K}}\} \subseteq \{1, \ldots, \hat{K}\}$ such that $|f_{i_1} - \hat{f}_{j_1}|, \ldots, |f_{i_{\mathcal{K}}} - \hat{f}_{j_{\mathcal{K}}}|$ are the smallest entries of the set $\{|f_i - \hat{f}_j| \ : \ 1 \leq i \leq K, 1 \leq j \leq \hat{K}\}$. For each $k \in \{1, \ldots, \mathcal{K}\}$, $\hat{f}_{j_k}$ is deemed to be the estimate for f_{i_k}. As usual, a Type I error is counted in connection with the frequency f_k if none of the estimated frequencies are assigned to f_k, and a Type II error is counted whenever $\hat{K} > K$. We compute also the mean-squared errors (MSE) for the frequency estimates.

For brevity, we report in Tables 6-9 the Type I and Type II errors together with the MSE for one single experiment conducted in each Example. In our comparisons, we consider SCp and the asymptotic criteria BIC and GAIC.

Because in Example 2 the simulated noise is an autoregressive process, we propose to analyze more carefully the data shown in Table 7. Remark that only GAIC has difficulties in recovering the first harmonic when $N > 35$, and recovering the second harmonic whose local SNR is smaller

SNR$_1$=1.00 dB								
N	30	40	50	60	70	80	90	100
SCp	**64**	**85**	**85**	**91**	**78**	**88**	**88**	**85**
SCa	51	58	61	69	67	80	82	81
SCe	59	75	83	**91**	77	85	**88**	**85**
BIC	32	55	58	73	70	78	70	78
KICc	60	69	65	75	59	58	58	59
GAIC	11	39	45	62	63	72	63	73
SNR$_1$=3.00 dB								
N	30	35	40	45	50	60	80	100
SCp	**93**	**86**	**94**	**96**	**92**	**94**	**93**	87
SCa	60	60	60	61	67	71	84	86
SCe	85	74	81	91	89	**94**	**93**	87
BIC	56	55	72	80	82	85	89	**91**
KICc	82	74	75	76	77	72	62	57
GAIC	50	60	70	84	80	84	90	89
SNR$_1$=5.00 dB								
N	25	30	35	40	45	50	75	100
SCp	**97**	**96**	**95**	**97**	**94**	**96**	92	94
SCa	61	58	61	68	61	59	82	89
SCe	83	86	85	86	88	**96**	92	94
BIC	53	71	72	85	80	86	93	93
KICc	90	89	81	88	79	81	63	59
GAIC	45	75	85	95	93	93	**95**	**95**

Table 4: Example 3: the performances in estimating the number of sine-waves reported with the same conventions as in Table 2.

posses problems to all the criteria. Note for SCp that the number of Type I errors connected with f_2 decreases fast with the increase of the sample size. For GAIC, the number of Type II errors is always small, but many Type I errors occur even for $N = 60$. This is a clear sign that, for small N, GAIC underestimates the number of sinusoids. The computed MSE is almost the same for all the investigated criteria and this is natural because the evaluation of SCp, BIC and GAIC is based on the estimates provided by the RELAX algorithm.

Final remarks

The new results on SC for the sinusoidal regression model illustrate very nicely the main idea that SC is not just the minus maximum log-likelihood term penalized with $\frac{k}{2}\ln N$, where k is the number of parameters and N is the number of samples. The most important achievement is to show that, for small and moderate sample sizes, the adequate use of SC could improve the estimation performances even for problems that have been intensively researched in the past, as

SNR$_1$=-5.00 dB								
N	30	40	50	60	70	80	90	100
SCp	24	65	**81**	**74**	**70**	**70**	**81**	84
SCa	22	65	**81**	**74**	**70**	**70**	**81**	**85**
SCe	32	64	70	66	63	66	79	83
BIC	15	53	79	64	65	69	70	79
KICc	**49**	**75**	79	70	**70**	67	73	71
GAIC	0	3	19	39	49	64	64	72
SNR$_1$=-3.00 dB								
N	30	40	50	60	70	80	90	100
SCp	20	82	**88**	**86**	**90**	**93**	**91**	**95**
SCa	20	**83**	**88**	**86**	**90**	91	84	89
SCe	33	71	80	78	83	87	86	92
BIC	13	72	**88**	79	78	83	90	94
KICc	**51**	**83**	84	81	80	78	74	72
GAIC	0	23	48	56	66	70	75	81
SNR$_1$=-1.00 dB								
N	30	35	40	45	50	60	80	100
SCp	30	80	**90**	**93**	**99**	**90**	**94**	92
SCa	30	80	86	**93**	97	88	77	73
SCe	40	72	74	80	90	83	91	87
BIC	27	83	85	87	95	**90**	92	95
KICc	**61**	**91**	89	89	90	82	76	65
GAIC	0	13	38	68	79	74	90	**97**

Table 5: Example 4: the performances in estimating the number of sine-waves reported with the same conventions as in Table 2.

it is the case with the mixed-spectrum estimation.

Acknowledgements

This work was supported by the Academy of Finland, project No. 113572, 118355 and 213462. The author is thankful to Gabriel Dospinescu from Ecole Normale Superieure, Paris, for the help with the proof of the Remark 2 in the Appendix.

Freq.	N		30	35	40	45	50	60	80	100
f_1	Err.1	SCp	0	0	0	0	0	0	0	0
		BIC	19	11	8	4	4	3	0	0
		GAIC	73	48	25	4	4	0	0	0
	MSE	SCp	-55.92	-56.45	-57.59	-60.01	-59.30	-62.62	-65.32	-68.79
		BIC	-56.15	-56.37	-57.52	-60.01	-59.17	-62.65	-65.32	-68.79
		GAIC	-56.61	-56.60	-58.30	-59.59	-59.04	-63.12	-65.32	-68.79
f_2	Err.1	SCp	38	25	10	7	5	0	0	0
		BIC	51	40	26	21	20	8	1	0
		GAIC	93	81	60	60	47	27	7	0
	MSE	SCp	-49.04	-51.43	-52.38	-54.00	-54.82	-56.64	-59.96	-63.19
		BIC	-46.09	-51.59	-51.90	-53.67	-54.47	-56.67	-59.92	-63.19
		GAIC	-46.08	-52.16	-51.71	-54.58	-55.88	-56.15	-59.97	-63.19
Err.2		SCp	0	3	2	0	0	1	2	1
		BIC	6	4	5	1	0	1	0	0
		GAIC	0	0	0	0	0	0	0	0

Table 6: Type I and Type II errors for Example 1 when SNR_2 = -1.00 dB. MSE is computed for the estimates of the frequencies and it is expressed in dB. The results are reported for 100 runs.

Freq.	N		30	35	40	45	50	60	80	100
f_1	Err.1	SCp	5	1	0	0	0	0	0	0
		BIC	9	8	1	0	0	1	0	0
		GAIC	28	20	14	12	9	9	1	0
	MSE	SCp	-58.96	-60.13	-61.20	-63.05	-64.48	-65.73	-71.68	-80.15
		BIC	-59.54	-59.87	-61.72	-63.53	-64.04	-65.69	-71.68	-80.15
		GAIC	-59.20	-60.62	-61.99	-63.66	-64.35	-65.66	-71.10	-80.15
f_2	Err.1	SCp	32	21	14	1	2	1	0	0
		BIC	24	29	20	7	7	4	1	0
		GAIC	83	74	66	61	42	22	1	0
	MSE	SCp	-49.97	-53.42	-55.07	-56.00	-42.96	-59.76	-64.03	-64.98
		BIC	-40.71	-39.15	-55.34	-56.82	-42.73	-60.11	-64.09	-64.98
		GAIC	-50.24	-53.27	-54.20	-56.21	-58.35	-59.84	-64.09	-64.98
Err.2		SCp	18	18	10	3	6	4	4	3
		BIC	54	30	27	22	16	5	4	2
		GAIC	10	9	6	1	4	3	2	2

Table 7: Type I and Type II errors for Example 2 when SNR_2 = 3.00 dB.

Freq.	N		30	35	40	45	50	60	80	100
f_1	Err.1	SCp	0	1	0	0	1	0	0	0
		BIC	5	8	1	2	3	1	0	2
		GAIC	43	32	20	14	18	9	4	6
	MSE	SCp	-52.69	-55.06	-56.26	-58.26	-59.12	-60.96	-64.09	-66.39
		BIC	-53.15	-54.38	-56.05	-58.34	-59.23	-60.97	-64.09	-66.50
		GAIC	-52.54	-55.38	-56.47	-58.01	-59.51	-60.93	-64.02	-66.52
f_2	Err.1	SCp	2	3	2	0	3	1	4	5
		BIC	20	21	16	11	13	11	10	7
		GAIC	50	40	27	16	19	15	10	11
	MSE	SCp	-53.61	-54.90	-58.91	-57.97	-60.02	-62.41	-65.47	-69.76
		BIC	-53.43	-56.14	-59.50	-58.48	-60.90	-62.32	-65.48	-69.67
		GAIC	-53.23	-54.98	-59.15	-57.59	-60.48	-62.21	-65.70	-69.87
Err.2		SCp	5	11	4	4	5	5	3	8
		BIC	24	24	12	9	5	4	1	2
		GAIC	0	0	3	0	1	1	0	0

Table 8: Type I and Type II errors for Example 3 when $SNR_1 = 3.00$ dB.

Freq.	N		30	40	50	60	70	80	90	100
f_1	Err.1	SCp	79	17	5	8	3	2	2	1
		BIC	85	23	8	18	17	13	6	4
		GAIC	100	77	52	43	34	29	22	17
	MSE	SCp	-42.86	-45.75	-51.69	-54.32	-55.19	-57.47	-58.26	-60.92
		BIC	-42.75	-46.07	-51.54	-47.16	-55.20	-57.33	-58.22	-60.76
		GAIC	-	-47.81	-52.23	-54.07	-56.13	-46.15	-58.84	-61.14
f_2	Err.1	SCp	79	6	1	0	0	0	0	0
		BIC	84	8	3	1	0	0	0	0
		GAIC	100	62	46	17	4	1	1	0
	MSE	SCp	-48.32	-55.79	-54.60	-59.31	-65.52	-65.26	-65.58	-65.73
		BIC	-50.14	-56.16	-54.60	-59.28	-65.49	-65.24	-65.58	-65.73
		GAIC	-	-55.54	-55.39	-59.32	-65.41	-65.24	-65.59	-65.73
f_3	Err.1	SCp	79	6	1	0	0	0	0	0
		BIC	84	8	3	1	0	0	0	0
		GAIC	100	62	46	17	4	1	1	0
	MSE	SCp	-45.49	-61.30	-58.20	-60.70	-68.24	-70.78	-82.14	-82.14
		BIC	-46.99	-61.45	-58.25	-60.31	-68.24	-70.78	-82.14	-82.14
		GAIC	-	-60.73	-59.16	-60.13	-68.40	-70.90	-82.14	-82.14
Err.2		SCp	0	1	7	6	7	5	7	4
		BIC	0	5	4	3	5	4	4	2
		GAIC	0	0	0	1	0	1	3	2

Table 9: Type I and Type II errors for Example 4 when $SNR_1 = -3.00$ dB.

APPENDIX

On the derivation of SC formula (10)

To check the conditions for the applicability of the SC formula of Qian and Künsch in our particular case, we resort to the closed-form expression of $\mathbf{J}_N(\boldsymbol{\theta})$ from the equations (4)-(7). We list below the conditions as they are given in [7]:

C1. $\mathbf{J}_N(\boldsymbol{\theta})$ is positive definite.

It is easy to check that all the eigenvalues of $\mathbf{J}_N(\boldsymbol{\xi}_k)$ are strictly positive. The covariance matrix $\mathbf{R}(\mathfrak{a})$ is positive definite for any M [30], therefore $\mathbf{J}_N(\mathfrak{a},\tau)$ is also positive definite, and the condition C1 is verified.

C2. The minimum eigenvalue of $\mathbf{J}_N(\boldsymbol{\theta})$ is of order $O(N)$ as $N \to \infty$.

Two of the eigenvalues of $\mathbf{J}_N(\boldsymbol{\xi}_k)$ are $O(N)$ and the third one is $O(N^3)$. As each eigenvalue of $\mathbf{J}_N(\mathfrak{a},\tau)$ is $O(N)$, we conclude that C2 is satisfied.

C3. $|\mathbf{J}_N(\boldsymbol{\theta}_1)|^{-1}||\mathbf{J}_N(\boldsymbol{\theta}_1)| - |\mathbf{J}_N(\boldsymbol{\theta}_2)|| \leq c||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||$, $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, where c is a finite constant.

For any $\boldsymbol{\theta}$, we have

$$|\mathbf{J}_N(\boldsymbol{\theta})| = N^{5K+M+1}\frac{|\mathbf{R}(\mathfrak{a})|}{2\tau^{M+2}}\prod_{k=1}^{K}|\mathbf{G}(\boldsymbol{\xi}_k,\mathfrak{a})|, \tag{15}$$

which implies that the left-hand-side term in the inequality C3 is finite and it does not depend on N. As the condition C3 is easily verified for $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, we analyze only the case $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. Thus we have $\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| > \delta$, where δ is given by the precision used to store the values of the parameters. To circumvent some technical difficulties, we consider firstly one sine-wave ($K = 1$) in white noise ($M = 0$). Without loss of generality, we assume $0 < \alpha_{min} < \alpha_1 < \alpha_{max} < \infty$ and $0 < \tau_{min} < \tau < \tau_{max} < \infty$. Elementary calculations lead to the inequality $\max_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}|\mathbf{J}_N(\boldsymbol{\theta}_1)|^{-1}||\mathbf{J}_N(\boldsymbol{\theta}_1)| - |\mathbf{J}_N(\boldsymbol{\theta}_2)|| < \Delta$, where $\Delta = (\alpha_{max}/\alpha_{min})^4(\tau_{max}/\tau_{min})^5$. Therefore, condition C3 is verified by selecting $c = \Delta/\delta$. To gain more insight, we assume next $K = 1$ and $M = 1$. As the noise model is stable, the AR coefficient is a non-zero number from the interval $(-1, 1)$. If supplementarily, the precision δ is used to store the value of the AR coefficient, then we get immediately $a_1 \in [-1+\delta, -\delta]\bigcup[\delta, 1-\delta]$. Taking α_1 and τ to be bounded as in the white noise case, it is not difficult to show that $\max_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}|\mathbf{J}_N(\boldsymbol{\theta}_1)|^{-1}||\mathbf{J}_N(\boldsymbol{\theta}_1)| - |\mathbf{J}_N(\boldsymbol{\theta}_2)|| < \Upsilon$, where $\Upsilon = (\alpha_{max}/\alpha_{min})^4(\tau_{max}/\tau_{min})^5((2-\delta)/\delta)^6$. Since $\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| > \upsilon$, we choose $c = \Upsilon/\upsilon$ and the condition C3 is verified. We emphasize that the precision used in this proof for the model parameters does not depend on the number of samples N.

C4. $\ln|\mathbf{J}_N(\boldsymbol{\theta})| = o(N)$.

Using the expression (15) for $|\mathbf{J}_N(\boldsymbol{\theta})|$, we readily obtain $\lim_{N\to\infty}\frac{\ln|\mathbf{J}_N(\boldsymbol{\theta})|}{N} = 0$, thus C4 is verified.

We apply next the SC criterion from [7]. For simplicity we ignore the terms that do not depend on N, and the SC formula becomes:

$$
\begin{aligned}
-\log f(y^N;\hat{\boldsymbol{\theta}}) + \log|\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)|^{1/2} + \sum_{i=1}^{3K+M+1}\log(|\hat{\theta}_i| + N^{-1/4}) \\
+ \sum_{i=1}^{3K+M+1} r^*(N^{1/4}|\hat{\theta}_i| + 1) + O(N^{-1/4}),
\end{aligned}
\tag{16}
$$

where $\log(\cdot)$ is the logarithm base 2, $\hat{\boldsymbol{\theta}}$ denotes the ML estimates, and $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N) = -\frac{\partial^2 \ln f(y^N;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ is the observed FIM. For any $x > 0$, $r^*(x) = \log(\log x) + \log(\log(\log x)) + \cdots$, where the sum continues as long as the iterated logarithms are strictly positive. The approximative formula (10) is obtained from (16) after operating the following changes:

- $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)$ is replaced with $\mathbf{J}_N(\hat{\boldsymbol{\theta}})$.
- An $O\left((3K+M+1)\log\log N\right)$ term is discarded.
- $\log(\cdot)$ is replaced with $\ln(\cdot)$.

Remark 1 The two-step encoding procedure adopted in [7] employs first a uniform quantization of Θ that is performed with the same precision for all the parameters. The term $N^{-1/4}$ in (10) is due to the option from [7] to select this precision based on the minimum eigenvalue of FIM.
Remark 2 It is recommended in [7] to consider in the SC expression also the term given by the number of parameters divided by two and multiplied by $\log\rho$, where ρ is the largest eigenvalue of $\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}$. We prove below that, under mild conditions, ρ does not depend on N, hence we ignore the $\log\rho$ term in (10).

Inspired by the expression of the asymptotic FIM, we assume there exist the non-singular matrices $\mathbf{A}$ and $\mathbf{B}$, and the diagonal matrix $\mathbf{C}_N$ such that $\mathbf{J}_N(\hat{\boldsymbol{\theta}}) = \mathbf{C}_N\mathbf{A}\mathbf{C}_N$ and $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N) = \mathbf{C}_N\mathbf{B}\mathbf{C}_N$. Supplementarily all the diagonal entries of $\mathbf{C}_N$ are powers of N, and the entries of $\mathbf{A}$ and $\mathbf{B}$ do not depend on N. With the notation $\mathbf{Z} = \mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}$, we have $\mathbf{Z} = \mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1/2}\left(\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1}\right)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{1/2}$, thus $\mathbf{Z}$ and $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\theta})^{-1}$ are similar. Moreover, $\tilde{\mathbf{J}}_N(\hat{\boldsymbol{\theta}},y^N)\mathbf{J}_N(\hat{\boldsymbol{\theta}})^{-1} = \mathbf{C}_N\mathbf{B}\mathbf{A}^{-1}\mathbf{C}_N^{-1}$, which leads to the conclusion that $\mathbf{Z}$ and $\mathbf{B}\mathbf{A}^{-1}$ are also similar. As the eigenvalues of $\mathbf{B}\mathbf{A}^{-1}$ do not depend on N, ρ is also independent of N.

References

[1] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory.* Prentice Hall, 1993.

[2] M. Ghogho and A. Swami, "Fast computation of the exact FIM for deterministic signals in colored noise," *IEEE Trans. Signal. Proces.*, vol. 47, no. 1, pp. 52–61, Jan. 1999.

[3] L. Kavalieris and E. Hannan, "Determining the number of terms in a trigonometric regression," *Journal of time series analysis*, vol. 15, no. 6, pp. 613–625, 1994.

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

[5] P. Stoica and A. Nehorai, "Statistical analysis of two nonlinear least-squares estimators of sine-wave parameters in the colored-noise case," *Circuits, Systems, and Signal Processing*, vol. 8, no. 1, pp. 3–15, 1989.

[6] P. Stoica and Y. Selen, "A review of information criterion rules," *IEEE Signal. Proces. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[7] G. Qian and H. Künsch, "Some notes on Rissanen's stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 782–786, Mar. 1998.

[8] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40–47, Jan. 1996.

[9] ——, "MDL denoising," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2537–2543, Nov. 2000.

[10] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.

[11] J. Rissanen, *Information and complexity in statistical modeling.* Springer Verlag, 2007.

[12] B. Friedlander and B. Porat, "The exact Cramer-Rao bound for Gaussian autoregressive processes," *IEEE Tr. on Aerospace and Electronic Systems*, vol. AES-25, pp. 3–8, 1989.

[13] C. Giurcăneanu and J. Rissanen, "Estimation of AR and ARMA models by stochastic complexity," in *Time series and related topics.*, H.-C. Ho, C.-K. Ing, and T. L. Lai, Eds. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 2006, vol. 52, pp. 48–59.

[14] C. Giurcăneanu, "Stochastic complexity for the estimation of sine-waves in colored noise," in *Proc. of 2007 Int. Conf. on Acoustics, Speech, and Signal Processing.* Honolulu, Hawaii, USA: IEEE, Apr. 2007, vol. 3, pp. 1097–1100.

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

[16] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

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

[18] B. G. Quinn, "Estimating the number of terms in a sinusoidal regression," *Journal of time series analysis*, vol. 10, no. 1, pp. 70–75, 1989.

[19] S. Kay, "Conditional model order estimation," *IEEE Trans. Signal. Proces.*, vol. 49, no. 9, pp. 1910–1917, Sep. 2001.

[20] ——, “Exponentially embedded families-new approaches to model order estimation,” *IEEE Trans. on Aerospace and Electronic Systems*, vol. 41, no. 1, pp. 333–345, Jan. 2005.

[21] A. Neath and J. Cavanaugh, “Regression and time series model selection using variants of the Schwarz information criterion,” *Communications in Statistics - Theory and Methods*, vol. 26, pp. 559–580, 1997.

[22] A. Hanson and P. C.-W. Fu, “Aplications of MDL to selected families of models,” in *Advances in Minimum Description Length: theory and applications*, P. Grünwald, I. Myung, and M. Pitt, Eds. MIT Press, 2005, ch. 5, pp. 125–150.

[23] E. Liski, “Normalized ML and the MDL principle for variable selection in linear regression,” in *Festschrift for Tarmo Pukkila on his 60th birthday*, E. Liski, J. Isotalo, J. Niemelä, S. Puntanen, and G. Styan, Eds. Univ. of Tampere, 2006, pp. 159–172.

[24] M. Hansen and B. Yu, “Minimum description length model selection criteria for generalized linear models,” in *Science and statistics: a festchrift for Terry Speed*, D. Goldstein, Ed. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 2002, vol. 40, pp. 145–164.

[25] E. George and D. Foster, “Calibration and empirical Bayesian variable selection,” *Biometrika*, vol. 87, no. 4, pp. 731–747, 2000.

[26] S. Bruzzone and M. Kaveh, “Information tradeoffs in using the sample autocorrelation function in ARMA parameter estimation,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 4, pp. 701–715, Aug. 1984.

[27] A.-K. Seghouane and M. Bekara, “A small sample model selection criterion based on Kullback’s symmetric divergence,” *IEEE Trans. Signal. Proces.*, vol. 52, pp. 3314–3323, 2004.

[28] M. Bekara, L. Knockaert, A.-K. Seghouane, and G. Fleury, “A model selection approach to signal denoising using Kullback’s symmetric divergence,” *Signal Processing*, vol. 86, pp. 1400–1409, 2006.

[29] J. Bernardo, “Psi (digamma) function,” *Appl. Statist.*, vol. 25, pp. 315–317, 1976.

[30] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. New Jersey: Prentice-Hall, 1997.