NEW INSIGHTS ON STOCHASTIC COMPLEXITY

Ciprian Doru Giurcăneanu and Seyed Alireza Razavi

Department of Signal Processing,
Tampere University of Technology, Finland
ciprian.giurcaneanu@tut.fi, alireza.razavi@tut.fi

ABSTRACT

The Minimum Description Length (MDL) principle led to various expressions of the stochastic complexity (SC), and the most recent one is given by the negative logarithm of the Normalized Maximum Likelihood (NML). For better understanding the properties of the newest SC-formula, we relate it to the well-known Generalized Likelihood Ratio Test (GLRT). Additionally, we compare the SC with the Bayesian Information Criterion (BIC) and other model selection rules. Some of the results are discussed in connection with families of models that are widely used in signal processing.

1. INTRODUCTION

The recent advances in model selection methods based on Minimum Description Length (MDL) principle have made possible to use the negative logarithm of the normalized maximum likelihood (NML) for evaluating the stochastic complexity (SC) [15]. The newest SC-formula (see Section 2 for more details) is not very popular in the signal processing community, where the asymptotic two-term criteria, called also GIC (General Information Criterion), are widely used. For GIC, 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, in some cases, on the sample size [1, 12, 17].

The two-term criteria have been extensively studied in the previous signal processing literature (see the survey paper [20]). Because it was introduced only recently, SC has received less attention. For example, the equivalence between GIC and the Generalized Likelihood Ratio Test (GLRT) was carefully investigated [19, 21], but similar results concerning the SC do not exist. This is why we elaborate in Section 3 on the relation between SC and GLRT. We mention that Rissanen has already used an early version of the SC for hypothesis testing [13]. Relying on the concept of optimally distinguishable distributions (ODD), an MDL method for composite hypothesis testing was proposed in [15] and further extended in [11]. In this paper we do not consider the ODD-based methods.

Section 4 is focused on the comparison between SC and other model selection rules, with a special emphasis on the relation between SC and the Bayesian Information Criterion (BIC) [17].

Notations: We denote vectors by boldface lowercase letters and matrices by boldface uppercase letters. The identity matrix of appropriate dimension is denoted by $\mathbf{I}$, while $\mathbf{0}$ denotes a null vector/matrix of appropriate dimension. For a full-rank matrix $\mathbf{X}$ having more rows than columns, $\langle\mathbf{X}\rangle$ is the column space of $\mathbf{X}$ and $\mathbf{P_X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the orthogonal projection onto $\langle\mathbf{X}\rangle$, with the convention that the superscripts $(\cdot)^\top$, $(\cdot)^{-1}$ denote the transpose and the matrix inverse, respectively. The projection onto the orthogonal complement of $\langle\mathbf{X}\rangle$ is $\mathbf{P}^\perp_{\mathbf{X}} = \mathbf{I} - \mathbf{P_X}$, while $\|\cdot\|$ is used for the Euclidean norm. The operator $|\cdot|$ denotes the determinant if the argument is a matrix, or the cardinality if the argument is a finite set. For an arbitrary $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the largest integer less than or equal to x. $\mathrm{B}(\cdot,\cdot)$ and $\Gamma(\cdot)$ are the Euler integrals of the first kind and the second kind, respectively. Notations for probability density functions: $\mathcal{N}(\mu,\mathbf{R})$ - the multivariate Gaussian distribution with mean μ and covariance matrix $\mathbf{R}$, χ^2_ν - the (central) chi-squared distribution with ν degrees of freedom, $\chi'^2_\nu(\lambda)$ - the non-central chi-squared distribution with ν degrees of freedom and noncentrality parameter λ. We denote $F'_{\nu_1,\nu_2}(\lambda_1)$ the (singly) non-central F-distribution, and we write $F'_{\nu_1,\nu_2}(\lambda_1) = [\chi'^2_{\nu_1}(\lambda_1)/\nu_1]/[\chi^2_{\nu_2}/\nu_2]$ to indicate that it is the ratio of a non-central chi-squared random variable and a central chi-squared random variable which are mutually independent. We also write $F''_{\nu_1,\nu_2}(\lambda_1,\lambda_2) = [\chi'^2_{\nu_1}(\lambda_1)/\nu_1]/[\chi^2_{\nu_2}(\lambda_2)/\nu_2]$ for the doubly non-central F-distribution, which is the ratio of two non-central chi-squared random variables that are mutually independent. It is obvious that $F''_{\nu_1,\nu_2}(\lambda_1,0) = F'_{\nu_1,\nu_2}(\lambda_1)$.

2. SC REVISITED

We focus on the SC formulation in the fundamental case of linear least squares regression problem, for which the observations $\mathbf{y}$ are modeled by:

$$\mathbf{y} = \mathbf{X}\theta + \varepsilon, \tag{1}$$

where $\mathbf{y} = [y_0,\ldots,y_{n-1}]^\top$, $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the regressor matrix having more rows than columns, $\theta \in \mathbb{R}^{m\times 1}$ is the vector of unknown parameters, and $\varepsilon \in \mathbb{R}^{n\times 1}$ is a vector whose entries are samples from an independent and identically distributed (i.i.d.) Gaussian process of zero mean and variance τ.

Because in most of the practical applications, not all the parameters $\theta_1,\ldots,\theta_m$ are equally important in modeling $\mathbf{y}$, one wants to eliminate those that are deemed to be unimportant. This reduces to choose a subset of the regressor variables indexed by $\gamma \subseteq M$, where $M = \{1,\ldots,m\}$. In line with the MDL principle, γ is selected such that to minimize the SC of the data $\mathbf{y}$ [15]. To give the SC-formula, we need some preparations. Let $\theta_\gamma \in \mathbb{R}^{|\gamma|\times 1}$ be the vector of the unknown regression coefficients within the γ-subset. The matrix $\mathbf{X}_\gamma$ is given by the columns of $\mathbf{X}$ that correspond to the γ-subset, and it is assumed to have full-rank. Similarly with (1), we have:

$$\mathbf{y} = \mathbf{X}_\gamma\theta_\gamma + \varepsilon_\gamma, \tag{2}$$

where the entries of ε_γ are i.i.d. Gaussian distributed with zero-mean and variance τ_γ. The noise variance τ_γ is unknown. The maximum likelihood (ML) estimate of θ_γ is $\hat\theta_\gamma = \left(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma\right)^{-1}\mathbf{X}_\gamma^\top\mathbf{y}$, which yields the residual sum of squares, $\mathrm{RSS}_\gamma = \|\mathbf{y} - \mathbf{X}_\gamma\hat\theta_\gamma\|^2$. Additionally, the ML estimate of the noise variance is $\hat\tau_\gamma = \mathrm{RSS}_\gamma/n$. When $|\gamma| > 0$, the stochastic complexity is evaluated as:

$$\begin{aligned}\mathrm{SC}(\mathbf{y};\gamma) \quad = \quad & (n-|\gamma|)\ln\hat\tau_\gamma + |\gamma|\ln\hat R_\gamma - 2\ln\Gamma\left(\frac{n-|\gamma|}{2}\right)\\ & -2\ln\Gamma\left(\frac{|\gamma|}{2}\right) + n\ln(n\pi) + 2L(\gamma) + 4\ln\ln\frac{b}{a},\end{aligned} \tag{3}$$

where

$$\hat R_\gamma = \|\mathbf{X}_\gamma\hat\theta_\gamma\|^2/n = \mathbf{y}^\top\mathbf{P}_{\mathbf{X}_\gamma}\mathbf{y}/n, \tag{4}$$

This work was supported by the Academy of Finland, project No. 113572, 118355 and 213462.

and $L(\gamma) = \min\left\{m, \left[\ln\binom{m}{|\gamma|} + \ln|\gamma| + \log_2(1+\ln m)\right]\right\}$ is the description length for the γ-structure. The real-valued hyper-parameters $b > a$ are generally neglected when comparing the SC computed for two different structures γ_1 and γ_2.

The derivation of SC was originally done in [14] and further refined in [15]. A similar criterion was derived in [4]. The expression in (3) is obtained after multiplying by two the SC formula from [15].

When $\gamma = \emptyset$, or equivalently, the observations $\mathbf{y}$ are assumed to be pure Gaussian noise with zero-mean and unknown variance τ_0, the stochastic complexity is given by:

$$\mathrm{SC}(\mathbf{y};\emptyset) = n\ln\hat{\tau}_0 - 2\ln\Gamma\left(\frac{n}{2}\right) + n\ln(n\pi) + 2\ln\ln\frac{b}{a}, \qquad (5)$$

where $\hat{\tau}_0 = \mathbf{y}^\top\mathbf{y}/n$ and the hyper-parameters a and b are the same as in (3). Similarly with (3), the equation in (5) is obtained after multiplying by two the corresponding formula from [15].

Selection of the best structure amounts to evaluate $\mathrm{SC}(\mathbf{y};\emptyset)$, then to calculate $\mathrm{SC}(\mathbf{y};\gamma)$ for all $\gamma \subseteq M$, and eventually to pick-up the subset that minimizes the stochastic complexity. To circumvent some computational difficulties, the previous literature recommends a simplified form of the criterion (3):

$$\mathrm{SC}(\mathbf{y};\gamma) = (n-|\gamma|)\ln\frac{\hat{\tau}_\gamma}{n-|\gamma|} + |\gamma|\ln\frac{\hat{R}_\gamma}{|\gamma|} + \ln[|\gamma|(n-|\gamma|)]. \qquad (6)$$

The formula above can be obtained from (3) in two steps: (i) neglect the sum $n\ln(n\pi) + 2L(\gamma) + 4\ln\ln\frac{b}{a}$; (ii) apply the Stirling approximation for the $\Gamma(\cdot)$ function and then discard all terms that do not depend on the γ-structure. The details of the calculations can be found in [9, 15]. Based on (6), the following result holds true.

Lemma 2.1. *Let γ be a collection of $|\gamma| > 0$ indices from M. If $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\bar{\theta}, \bar{\tau}\mathbf{I})$, then*

$$\begin{aligned}\mathrm{SC}(\mathbf{y};\gamma) \quad &= \quad (n-|\gamma|)\ln\left[\frac{1}{n-|\gamma|}\chi'^2_{n-|\gamma|}\left(\lambda_\gamma^{(1)}\right)\right] \\ &\quad +|\gamma|\ln\left[\frac{1}{|\gamma|}\chi'^2_{|\gamma|}\left(\lambda_\gamma^{(2)}\right)\right] + \ln[|\gamma|(n-|\gamma|)] + ct, (7)\end{aligned}$$

where $\lambda_\gamma^{(1)} = \|\mathbf{P}^\perp_{\mathbf{X}_\gamma}\mathbf{X}\bar{\theta}\|^2/\bar{\tau}$, $\lambda_\gamma^{(2)} = \|\mathbf{P}_{\mathbf{X}_\gamma}\mathbf{X}\bar{\theta}\|^2/\bar{\tau}$, the two χ'^2's are mutually independent, and $ct = n\ln(\bar{\tau}/n)$.

See Appendix A.1 for the proof. □

Lemma 2.1 will be used in Section 4 to compare SC with other model selection criteria. In the next section, we investigate the relationship between SC and GLRT.

3. SC AND GLRT

For an arbitrary γ-structure, we note that the density function of the measurements $\mathbf{y}$ is:

$$f(\mathbf{y};\theta_\gamma,\tau_\gamma) = \frac{1}{(2\pi\tau_\gamma)^{n/2}}\exp\left(-\frac{1}{2\tau_\gamma}\|\mathbf{y}-\mathbf{X}_\gamma\theta_\gamma\|^2\right), \qquad (8)$$

as it can be easily observed from (2). To investigate the relation between SC and GLRT, we consider the problem of selecting between the model classes:

$$\begin{aligned}\mathcal{M}_0 &= \{f(\mathbf{y};\theta_0,\tau_0) : \theta_0 \in \mathbb{R}^{m\times 1}, \theta_0 = \mathbf{0}, \tau_0 > 0\}, \\ \mathcal{M}_1 &= \{f(\mathbf{y};\theta_1,\tau_1) : \theta_1 \in \mathbb{R}^{m\times 1}, \theta_1 \neq \mathbf{0}, \tau_1 > 0\}.\end{aligned}$$

For writing the equations more compactly, we take Λ to be the usual GLRT [6] multiplied by two, $\Lambda = 2\ln\frac{f(\mathbf{y};\hat{\theta}_1,\hat{\tau}_1)}{f(\mathbf{y};\mathbf{0},\hat{\tau}_0)}$, where $\hat{\theta}_1$ and $\hat{\tau}_1$ are the ML estimates for the model class $\mathcal{M}_1$, while $\hat{\tau}_0$ is the ML estimate for the model class $\mathcal{M}_0$. We have the following chain of identities

$$\begin{aligned}\Lambda &= n\ln\frac{\hat{\tau}_0}{\hat{\tau}_1} \qquad (9) \\ &= n\ln\frac{\mathbf{y}^\top\mathbf{y}/n}{\mathbf{y}^\top\mathbf{P}^\perp_{\mathbf{X}}\mathbf{y}/n} \\ &= n\ln\left(1+\frac{\mathbf{y}^\top\mathbf{P}_{\mathbf{X}}\mathbf{y}/n}{\mathbf{y}^\top\mathbf{P}^\perp_{\mathbf{X}}\mathbf{y}/n}\right) \\ &= n\ln\left(1+\frac{\hat{R}_1}{\hat{\tau}_1}\right). \qquad (10)\end{aligned}$$

In the calculations above we used the fact that $\mathbf{X}$ is the regressor matrix for $\mathcal{M}_1$ (see also (1)). The identity in (10) was obtained by applying the definition in (4) to compute $\hat{R}_1$ for the model class $\mathcal{M}_1$. Let $\gamma_0 = \emptyset$ and $\gamma_1 = M$. After some algebraic manipulations, the formulas in (3) and (5) lead to:

$$\begin{aligned}\mathrm{SC}(\mathbf{y};\gamma_1) - \mathrm{SC}(\mathbf{y};\gamma_0) \quad &= \quad n\ln\frac{\hat{\tau}_1}{\hat{\tau}_0} + m\ln\frac{\hat{R}_1}{\hat{\tau}_1} - 2\ln\mathrm{B}\left(\frac{m}{2},\frac{n-m}{2}\right) \\ &\quad +2L(\gamma_1) + 2\ln\ln\frac{b}{a}. \qquad (11)\end{aligned}$$

The first two terms in (11) can be expressed as a function of Λ via (9) and (10). So,

$$\begin{aligned}\mathrm{SC}(\mathbf{y};\gamma_1) - \mathrm{SC}(\mathbf{y};\gamma_0) \quad &= \quad \rho(m,n,\Lambda) + \eta(m,n) + 2\ln\ln\frac{b}{a}, (12) \\ \rho(m,n,\Lambda) \quad &= \quad -\Lambda + m\ln\left[\exp\left(\frac{\Lambda}{n}\right) - 1\right], \qquad (13) \\ \eta(m,n) \quad &= \quad 2\min\{m, [\ln m + \log_2(1+\ln m)]\} \\ &\quad -2\ln\mathrm{B}\left(\frac{m}{2},\frac{n-m}{2}\right). \qquad (14)\end{aligned}$$

This helps us to prove the following result.

Proposition 3.1. *For n and m fixed ($n > m+1$), we have:*
i) The difference $\mathrm{SC}(\mathbf{y};\gamma_1) - \mathrm{SC}(\mathbf{y};\gamma_0)$ is a concave function of Λ, which increases on $(-\infty,\Lambda^*)$ and decreases on $[\Lambda^*,\infty)$, where $\Lambda^* = n\ln[n/(n-m)]$.
ii) $\max_\Lambda \rho(m,n,\Lambda) + \eta(m,n) \leq 0$ if and only if $m = 1$.

See Appendix A.2 for the proof. □

Discussion The usual F-statistic for testing if the observations $\mathbf{y}$ are outcomes of the model class $\mathcal{M}_0$ has the expression $F = \frac{\hat{R}_1/m}{\hat{\tau}_1/(n-m)} = \frac{n-m}{m}\left[\exp\left(\frac{\Lambda}{n}\right) - 1\right]$ [18]. Remark that $\Lambda = \Lambda^*$ is equivalent with $F = 1$. Therefore, according to the point (i) of the Proposition 3.1, whenever $F > 1$, the selection rule based on SC will choose the model class $\mathcal{M}_1$ if and only if Λ is larger than a certain threshold. This shows clearly that SC is equivalent with GLRT when $F > 1$. The situation changes when $F \leq 1$: SC selects $\mathcal{M}_1$ if Λ is smaller than a certain threshold.
The second part of the Proposition 3.1 points out that, for $m = 1$, SC will always prefer the model class $\mathcal{M}_1$ to $\mathcal{M}_0$ if the term given by the hyper-parameters a and b is neglected. The drawback does not exist when $m > 1$, and this is probably the main reason for which, in the previous literature, the term $4\ln\ln\frac{b}{a}$ of the criterion (3) was neglected. We refer to [15] for more details on the computation of a and b.

In the next section, we extend our analysis to the relation between SC and BIC.

4. SC AND BIC

It is well-known that BIC has the expression [17],

$$\mathrm{BIC}(\mathbf{y};\gamma) = n\ln\hat{\tau}_\gamma + |\gamma|\ln n, \tag{15}$$

and is equivalent with the crude MDL criterion from [12]. To make easier the comparison between SC and BIC, we re-write the formula from (6) as:

$$\mathrm{SC}(\mathbf{y};\gamma) = n\ln\hat{\tau}_\gamma + |\gamma|\ln F_\gamma + \ln\left[|\gamma|/(n-|\gamma|)^{n-1}\right], \tag{16}$$

where $F_\gamma = \frac{\hat{R}_\gamma/|\gamma|}{\hat{\tau}_\gamma/(n-|\gamma|)}$. From (15) and (16), we observe that the goodness-of-fit term is the same for both criteria. The difference is given by the penalty term, which for SC has two components. The most intriguing is $|\gamma|\ln F_\gamma$, which depends not only on the number of parameters and the sample size, but also on the measurements $\mathbf{y}$. Remark that $|\gamma|\ln F_\gamma$ is strictly positive only when $F_\gamma > 1$.

In [4], it is given a theoretical result on the behavior of F_γ in the case when $m \to \infty$ and the entries of $\mathbf{X}$ are random. Based on this analysis, Hansen and Yu conclude that the model selection rules for which $|\gamma|\ln F_\gamma$ is a penalty term combine the strengths of both Akaike Information Criterion (AIC) [1] and BIC [17].

In most of the signal processing problems, m is assumed to be finite and the matrix $\mathbf{X}$ is deterministic. To compare SC and BIC under these assumptions, we compute $E[F_\gamma]$. It is clear that SC and BIC are about the same if $E[F_\gamma] = O(n)$. A similar fact was pointed out in [8] when comparing the Exponentially Embedded Families (EEF) criterion and the BIC.

We focus on the problem of selection between two nested models. Let us partition the full-rank matrix $\mathbf{X}$ from (1) into two blocks such that $\mathbf{X}_0$ contains the first m_0 columns of $\mathbf{X}$ and $\mathbf{X}_i$ is formed by the rest of the columns. With the convention that $0 < m_0 < m$, we can write $\mathbf{X} = [\mathbf{X}_0\ \mathbf{X}_i]$. Let $\gamma_0 = \{1,\dots,m_0\}$ and $\gamma_1 = M$. We consider the model classes:

$$\begin{aligned} \mathcal{M}_{\gamma_0} &= \{f(\mathbf{y};\theta_0,\tau_0) &:\ & \theta_0 \in \mathbb{R}^{m_0\times 1}\setminus\{\mathbf{0}\}, \tau_0 > 0\},\\ \mathcal{M}_{\gamma_1} &= \{f(\mathbf{y};[\theta_0^\top\ \theta_i^\top]^\top,\tau_1) &:\ & \theta_0 \in \mathbb{R}^{m_0\times 1}\setminus\{\mathbf{0}\},\\ & & & \theta_i \in \mathbb{R}^{(m-m_0)\times 1}\setminus\{\mathbf{0}\}, \tau_1 > 0\}, \end{aligned}$$

where $f(\mathbf{y};\theta_{\gamma_h},\tau_{\gamma_h})$ with $h \in \{0,1\}$ is the normal density function from (8).

With slight abuse of notation, we write $\mathbf{y} \in \mathcal{M}_{\gamma_0}$ if $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_0\theta_0,\tau_0)$. Similarly, $\mathbf{y} \in \mathcal{M}_{\gamma_1}$ means $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_0\theta_0 + \mathbf{X}_i\theta_i,\tau_1)$. Our goal is to compute $E[F_{\gamma_0}]$ and $E[F_{\gamma_1}]$ when $\mathbf{y} \in \mathcal{M}_{\gamma_0}$ or $\mathbf{y} \in \mathcal{M}_{\gamma_1}$. We formalize the result as follows:

Proposition 4.1. *i) For $h \in \{0,1\}$, we have:*

$$\begin{aligned} \mathrm{SC}(\mathbf{y};\gamma_h) &= n\ln\hat{\tau}_{\gamma_h} + |\gamma_h|\ln\left[F''_{|\gamma_h|,n-|\gamma_h|}(\lambda^{(2)}_{\gamma_h},\lambda^{(1)}_{\gamma_h})\right]\\ &\quad + \ln\left[|\gamma_h|/(n-|\gamma_h|)^{(n-1)}\right]. \end{aligned} \tag{17}$$

The values of $\lambda^{(1)}_{\gamma_h}$ and $\lambda^{(2)}_{\gamma_h}$ are given below.

	$\lambda^{(1)}_{\gamma_0}$	$\lambda^{(2)}_{\gamma_0}$	$\lambda^{(1)}_{\gamma_1}$	$\lambda^{(2)}_{\gamma_1}$
$\mathbf{y} \in \mathcal{M}_{\gamma_0}$	0	$\frac{\lVert\mathbf{X}_0\theta_0\rVert^2}{\tau_0}$	0	$\frac{\lVert\mathbf{X}_0\theta_0\rVert^2}{\tau_0}$
$\mathbf{y} \in \mathcal{M}_{\gamma_1}$	$\frac{\lVert\mathbf{P}^\perp_{\mathbf{X}_0}\mathbf{X}_i\theta_i\rVert^2}{\tau_1}$	$\frac{\lVert\mathbf{X}_0\theta_0+\mathbf{P}_{\mathbf{X}_0}\mathbf{X}_i\theta_i\rVert^2}{\tau_1}$	0	$\frac{\lVert\mathbf{X}_0\theta_0+\mathbf{X}_i\theta_i\rVert^2}{\tau_1}$

ii) If $n > m+2$, then:

$$E\left[F''_{|\gamma_0|,n-|\gamma_0|}(\lambda^{(2)}_{\gamma_0},\lambda^{(1)}_{\gamma_0})\right]$$

$$= \begin{cases} \left[1+\frac{2}{n-m_0-2}\right]\left[1+\frac{\|\mathbf{X}_0\theta_0\|^2/\tau_0}{m_0}\right], & \mathbf{y} \in \mathcal{M}_{\gamma_0},\\ \left[1+\frac{2}{n-m_0-2}\right]\frac{1+\frac{\|\mathbf{X}_0\theta_0+\mathbf{P}_{\mathbf{X}_0}\mathbf{X}_i\theta_i\|^2/\tau_1}{m_0}}{1+\frac{\|\mathbf{P}^\perp_{\mathbf{X}_0}\mathbf{X}_i\theta_i\|^2/\tau_1}{n-m_0}}, & \mathbf{y} \in \mathcal{M}_{\gamma_1}, \end{cases} \tag{18}$$

$$E\left[F''_{|\gamma_1|,n-|\gamma_1|}(\lambda^{(2)}_{\gamma_1},\lambda^{(1)}_{\gamma_1})\right]$$

$$= \begin{cases} \left[1+\frac{2}{n-m-2}\right]\left[1+\frac{\|\mathbf{X}_0\theta_0\|^2/\tau_0}{m}\right], & \mathbf{y} \in \mathcal{M}_{\gamma_0},\\ \left[1+\frac{2}{n-m-2}\right]\left[1+\frac{\|\mathbf{X}_0\theta_0+\mathbf{X}_i\theta_i\|^2/\tau_1}{m}\right], & \mathbf{y} \in \mathcal{M}_{\gamma_1}. \end{cases} \tag{19}$$

The value of $E\left[F''_{|\gamma_0|,n-|\gamma_0|}(\lambda^{(2)}_{\gamma_0},\lambda^{(1)}_{\gamma_0})\right]$ for $\mathbf{y} \in \mathcal{M}_{\gamma_1}$ is an approximation obtained under the condition:

$$\frac{\|\mathbf{P}^\perp_{\mathbf{X}_0}\mathbf{X}_i\theta_i\|^2/\tau_1}{n-m_0} < \frac{1}{2}, \tag{20}$$

whereas all other results in Proposition 4.1 are exact. The proof is deferred to Appendix A.3. □

Next we analyze how $E[F_\gamma]$ is affected by the structure of $\mathbf{X}$.

Case 1: $\mathbf{P}_{\mathbf{X}_0}\mathbf{X}_i \approx \mathbf{0}$. We consider the problem of estimating the number of sine-waves in white Gaussian noise, when the frequencies $\omega_1,\dots,\omega_{m/2} \in (0,\pi)$ are known, but the amplitudes $\alpha_1,\dots,\alpha_{m/2} > 0$ and the phases $\phi_1,\dots,\phi_{m/2} \in [-\pi,\pi)$ are unknown. Hence, the regressor matrix has the expression

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ \cos[\omega_1(n-1)] & \sin[\omega_1(n-1)] & \cdots & \sin[\omega_{m/2}(n-1)] \end{bmatrix},$$

$\theta_0 = [\alpha_1\cos\phi_1, -\alpha_1\sin\phi_1,\dots,-\alpha_{m_0/2}\sin\phi_{m_0/2}]^\top$ and $\theta_i = [\alpha_{m_0/2+1}\cos\phi_{m_0/2+1}, -\alpha_{m_0/2+1}\sin\phi_{m_0/2+1},\dots,-\alpha_{m/2}\sin\phi_{m/2}]^\top$. Without loss of generality, we assume $\tau_0 = \tau_1 = \tau$. It is well-known the definition of the local SNR for the k-th sinusoid: $\mathrm{SNR}_k = \alpha_k^2/(2\tau)$. We take $S_0 = \|\theta_0\|^2/(2\tau)$ to be the sum of the SNR's for all sine-waves when the model class is $\mathcal{M}_{\gamma_0}$. Similarly, $S_i = \|\theta_i\|^2/(2\tau)$. With the approximation $\mathbf{X}^\top\mathbf{X} \approx (n/2)\mathbf{I}$ (see [2, 8] for more details), the condition in (20) reduces to $[n/(n-m_0)]S_i < 1/2$, and the identities in (18) and (19) become:

$$E\left[F''_{|\gamma_0|,n-|\gamma_0|}(\lambda^{(2)}_{\gamma_0},\lambda^{(1)}_{\gamma_0})\right]$$

$$\approx \begin{cases} \left[1+\frac{2}{n-m_0-2}\right]\left[1+n\frac{S_0}{m_0}\right], & \mathbf{y} \in \mathcal{M}_{\gamma_0},\\ \left[1+\frac{2}{n-m_0-2}\right]\left[1+\frac{1+n(S_0/m_0)}{1+[n/(n-m_0)]S_i}\right], & \mathbf{y} \in \mathcal{M}_{\gamma_1}, \end{cases} \tag{21}$$

$$E\left[F''_{|\gamma_1|,n-|\gamma_1|}(\lambda^{(2)}_{\gamma_1},\lambda^{(1)}_{\gamma_1})\right]$$

$$\approx \begin{cases} \left[1+\frac{2}{n-m-2}\right]\left[1+n\frac{S_0}{m}\right], & \mathbf{y} \in \mathcal{M}_{\gamma_0},\\ \left[1+\frac{2}{n-m-2}\right]\left[1+n\frac{S_0+S_i}{m}\right], & \mathbf{y} \in \mathcal{M}_{\gamma_1}. \end{cases} \tag{22}$$

From (15), (16), (21) and (22), one can note that SC and BIC are about the same when m is fixed and $n \gg m$. More interestingly, when $\mathbf{y} \in \mathcal{M}_{\gamma_1}$, if m,n,m_0,S_0 are kept fixed and S_i increases such that $[n/(n-m_0)]S_i$ is not larger than $1/2$, then $E\left[F''_{|\gamma_0|,n-|\gamma_0|}(\lambda^{(2)}_{\gamma_0},\lambda^{(1)}_{\gamma_0})\right]$ decreases and $E\left[F''_{|\gamma_1|,n-|\gamma_1|}(\lambda^{(2)}_{\gamma_1},\lambda^{(1)}_{\gamma_1})\right]$ increases. The effect produced by the increase of S_i cannot be understood properly if we restrict our attention only to the penalty term of the SC criterion. By observing that $\ln\hat{\tau}_\gamma = -\ln\left(\frac{|\gamma|}{n-|\gamma|}F_\gamma + 1\right) + \ln\frac{\mathbf{y}^\top\mathbf{y}}{n}$, we get

$\frac{\partial \mathrm{SC}(\gamma)}{\partial F_\gamma} = \frac{|\gamma|(n-|\gamma|)}{F_\gamma(|\gamma|F_\gamma + n - |\gamma|)}(1 - F_\gamma)$. Provided that $F_\gamma > 1$, $\mathrm{SC}(\gamma)$ is a decreasing function of F_γ. It means that whenever S_i increases, $E\left[F''_{|\gamma_0|,n-|\gamma_0|}(\lambda^{(2)}_{\gamma_0}, \lambda^{(1)}_{\gamma_0})\right]$ decreases, which corresponds to an increase of $\mathrm{SC}(\gamma_0)$. Moreover, when S_i increases, $\mathrm{SC}(\gamma_1)$ decreases. In other words, an increase of the SNR of the k-th sine-wave when $m_0/2 < k \leq m/2$ will decrease the difference $\mathrm{SC}(\gamma_1) - \mathrm{SC}(\gamma_0)$ such that to favor the model class $\mathcal{M}_{\gamma_1}$.

Case 2: $\mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{X}_i \approx \mathbf{0}$. For simplicity, we assume $m = m_0 + 1$. Thus, $\mathbf{X}_i$ is the last column of $\mathbf{X}$, and we prefer to use the notation $\mathbf{x}_i$ instead of $\mathbf{X}_i$. In [8], it was pointed out that the Fisher information matrix (FIM) for the model class $\mathcal{M}_{\gamma_1}$ is badly conditioned numerically when $\langle \mathbf{X}_0 \rangle$ and $\langle \mathbf{X}_i \rangle$ are nearly colinear. Based on this observation, Kay concluded that, whenever $\mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{x}_i \approx \mathbf{0}$, the model selection rules whose penalty term is given by the determinant of the FIM [2, 7, 10] will always choose the most complex model. This drawback does not exist for SC. When $\langle \mathbf{X}_0 \rangle$ and $\langle \mathbf{X}_i \rangle$ are nearly colinear, the condition in (20) is satisfied, and from (18) and (19) we get for $\mathbf{y} \in \mathcal{M}_{\gamma_1}$:

$$E\left[F''_{|\gamma_0|,n-|\gamma_0|}(\lambda^{(2)}_{\gamma_0}, \lambda^{(1)}_{\gamma_0})\right] \approx \left[1 + \frac{2}{n-m-1}\right]\left[1 + \frac{\|\mathbf{X}\theta\|^2/\tau}{m-1}\right],$$

$$E\left[F''_{|\gamma_1|,n-|\gamma_1|}(\lambda^{(2)}_{\gamma_1}, \lambda^{(1)}_{\gamma_1})\right] \approx \left[1 + \frac{2}{n-m-2}\right]\left[1 + \frac{\|\mathbf{X}\theta\|^2/\tau}{m}\right],$$

where $\theta = [\theta_0^\top \ \theta_i^\top]^\top$. For $\mathbf{y} \in \mathcal{M}_{\gamma_0}$, the expressions of the expectation of F_γ are the same as in (18) and (19). From these results, it is evident that SC does not favor the $\mathcal{M}_{\gamma_1}$ class model even if $\mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{x}_i \approx \mathbf{0}$.

Case 3: the largest entry of $\mathbf{X}^\top\mathbf{X}$ is $O(n^\zeta)$ with $\zeta > 1$. Let us assume that both $\mathcal{M}_{\gamma_0}$ and $\mathcal{M}_{\gamma_1}$ model polynomials embedded in noise. The degree of the polynomial is $m_0 - 1$ for $\mathcal{M}_{\gamma_0}$ and $m - 1$ for $\mathcal{M}_{\gamma_1}$. Therefore, the largest entry of $\mathbf{X}^\top\mathbf{X}$ is $O(n^{2m-1})$ [2]. For simplicity, we take $m = m_0 + 1$. Note that $F_{\gamma_0} = O(n^{2m-3})$ and $F_{\gamma_1} = O(n^{2m-1})$. This shows that SC penalizes more stringently than BIC the polynomial models with higher degree, and explains why in the experimental results reported in Table 9.1 from [15] BIC overestimates the polynomial degree more often than SC.

Because $m = m_0 + 1$, we have $\mathbf{X}_i = \mathbf{x}_i$ like in the previous example, and for verifying the condition in (20) we must compute $\mathbf{x}_i^\top \mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{x}_i$. By using the identity $\mathbf{x}_i^\top \mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{x}_i = |\mathbf{X}^\top\mathbf{X}|/|\mathbf{X}_0^\top\mathbf{X}_0|$ from [8], together with the results from [2], we obtain $\mathbf{x}_i^\top \mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{x}_i = O(n^{2m-1})$. For m and n large, the condition in (20) is not satisfied and the approximation of $E\left[F''_{|\gamma_0|,n-|\gamma_0|}(\lambda^{(2)}_{\gamma_0}, \lambda^{(1)}_{\gamma_0})\right]$ given in (18) must be applied with caution.

5. CONCLUDING REMARKS

The analysis of the relationship between SC and GLRT shows the importance of the hyper-parameters within SC-formula. The comparison between SC and other information theoretic criteria reveals the robustness of SC for families of models commonly used in practice. A more detailed investigation of this aspect, along with a simulation study, will be the topic for a future research paper.

A. APPENDIX

A.1 Proof of Lemma 2.1

The proof is a straightforward consequence of the following result:

Theorem A.1. *[3] Let $\mathbf{y} \sim \mathcal{N}(\mu, \tau\mathbf{I})$, and let $\sum_{\ell=1}^{q} \mathbf{y}^\top\mathbf{B}_\ell\mathbf{y} = \mathbf{y}^\top\mathbf{y}$, where the rank of $\mathbf{B}_\ell$ is r_ℓ. Anyone of the three conditions listed below is a necessary and sufficient condition that the following two statements be true: (1) $\mathbf{y}^\top\mathbf{B}_\ell\mathbf{y}/\tau \sim \chi'^2_{r_\ell}(\lambda_\ell)$, where $\lambda_\ell = \mu^\top\mathbf{B}_\ell\mu/\tau$; (2) $\mathbf{y}^\top\mathbf{B}_\ell\mathbf{y}$ and $\mathbf{y}^\top\mathbf{B}_{\ell'}\mathbf{y}$ are independent if $\ell \neq \ell'$. The conditions are: (c1) $\mathbf{B}_\ell$ is idempotent for all $\ell \in \{1, \ldots, q\}$; (c2) $\mathbf{B}_\ell\mathbf{B}_{\ell'} = \mathbf{0}$ for all $\ell \neq \ell'$; (c3) $\sum_{\ell=1}^{q} r_\ell = n$.*

Note that $n\hat{\tau}_\gamma/\bar{\tau} = \mathbf{y}^\top\mathbf{P}^{\perp}_{\mathbf{X}_\gamma}\mathbf{y}/\bar{\tau}$ and $n\hat{R}_\gamma/\bar{\tau} = \mathbf{y}^\top\mathbf{P}_{\mathbf{X}_\gamma}\mathbf{y}/\bar{\tau}$. By applying the theorem above, we readily obtain $n\hat{\tau}_\gamma/\bar{\tau} \sim \chi'^2_{n-|\gamma|}\left(\|\mathbf{P}^{\perp}_{\mathbf{X}_\gamma}\mathbf{X}\bar{\theta}\|^2/\bar{\tau}\right)$ and $n\hat{R}_\gamma/\bar{\tau} \sim \chi'^2_{|\gamma|}\left(\|\mathbf{P}_{\mathbf{X}_\gamma}\mathbf{X}\bar{\theta}\|^2/\bar{\tau}\right)$. Next we use (6), and we take the term ct, which does not depend on γ, to be $n\ln(\bar{\tau}/n)$. Moreover, based on Theorem A.1, it is evident that the two χ'^2's involved in formula (7) are mutually independent. □

A.2 Proof of Proposition 3.1

i) It is easy to check via (12)-(14) that the difference $\mathrm{SC}(\mathbf{y};\gamma_1) - \mathrm{SC}(\mathbf{y};\gamma_0)$ is a concave function of Λ, and its first derivative is given by $\frac{m/n}{1-\exp(-\Lambda/n)} - 1$. The result on the monotonicity of $\mathrm{SC}(\mathbf{y};\gamma_1) - \mathrm{SC}(\mathbf{y};\gamma_0)$ is straightforward.
ii) By relying on the point (i) of the Proposition, it can be shown for all $m \in \{1, \ldots, n-2\}$ that:

$$\begin{aligned}
&\max_{\Lambda} \rho(m,n,\Lambda) + \eta(m,n) \\
&\quad = -n\ln\frac{n}{n-m} + m\ln\frac{m}{n-m} - 2\ln \mathrm{B}\left(\frac{m}{2}, \frac{n-m}{2}\right) \\
&\qquad +2\min\{m, [\ln m + \log_2(1+\ln m)]\}.
\end{aligned}$$

First we consider the case $m = 1$. After simple manipulations that exploit the properties of the $\mathrm{B}(\cdot,\cdot)$ and $\Gamma(\cdot)$ functions, we get:

$$\begin{aligned}
&\max_{\Lambda} \rho(1,n,\Lambda) + \eta(1,n) \\
&\quad = -n\ln\frac{n}{n-1} + \ln\frac{1}{n-1} - \ln\pi + 2\ln\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \\
&\quad < -n\ln\frac{n}{n-1} + \ln\frac{1}{n-1} - \ln\pi + \ln\frac{n-1}{2} \qquad (23) \\
&\quad = -n\ln\frac{n}{n-1} - \ln(2\pi) < 0.
\end{aligned}$$

The inequality in (23) was obtained by applying the well-known inequality: $\frac{\Gamma(x+1)}{\Gamma\left(x+\frac{1}{2}\right)} < \left(x + \frac{1}{2}\right)^{1/2}$ for all $x > -\frac{1}{4}$.

It remains to demonstrate that $\max_{\Lambda} \rho(m,n,\Lambda) + \eta(m,n) > 0$ when $m \geq 2$. For $2 \leq m < \lfloor n/2 \rfloor$, we have $\max_{\Lambda} \rho(m,n,\Lambda) = \max_{\Lambda} \rho(n-m,n,\Lambda)$ and $\eta(m,n) < \eta(n-m,n)$. Hence, it is enough to prove that $\max_{\Lambda} \rho(m,n,\Lambda) + \eta(m,n) > 0$ for $2 \leq m \leq \lfloor n/2 \rfloor$. When $m = 2$, simple calculations lead to $\max_{\Lambda} \rho(2,n,\Lambda) + \eta(2,n) = -n\ln\frac{n}{n-2} + 2\ln 2 + 2\log_2(1+\ln 2)$, which is positive for $n = 4$. Because the negative term $-n\ln\frac{n}{n-2}$ is an increasing function of n, we obtain immediately that $\max_{\Lambda} \rho(2,n,\Lambda) + \eta(2,n) > 0$ for all $n \geq 4$. A similar result holds for $m = 3$.

Next we show that switching from m to $m+2$ when n is kept constant and $1 < m < \lfloor n/2 \rfloor$, will have as effect an increase of the maximum possible value of the difference $\mathrm{SC}(\mathbf{y};\gamma_1) - \mathrm{SC}(\mathbf{y};\gamma_0)$. This can be done in two steps. First step is to observe that $L(\gamma)$ is an increasing function of m. Second step is to check the equality

$$\begin{aligned}
&\left[\max_{\Lambda} \rho(m+2,n,\Lambda) - 2\ln \mathrm{B}\left(\frac{m+2}{2}, \frac{n-m-2}{2}\right)\right] \\
&\qquad - \left[\max_{\Lambda} \rho(m,n,\Lambda) - 2\ln \mathrm{B}\left(\frac{m}{2}, \frac{n-m}{2}\right)\right] = g(m) - g(n-m-2),
\end{aligned}$$

where $g(x) = (x+2)\ln[(x+2)/x]$ for all $x > 0$. The fact that $g(x)$ is monotonically decreasing for $x \geq 1$ completes the proof.

□

A.3 Proof of Proposition 4.1

i) Lemma 2.1 together with (16) leads to (17). Let us define: $\mathbf{G} = \mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{X}_i$. The following two identities are known (see, for example, [16]): $\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{X}_0} + \mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{P}_{\mathbf{G}}\mathbf{P}^{\perp}_{\mathbf{X}_0}$ and $\mathbf{P}^{\perp}_{\mathbf{X}} = \mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{P}^{\perp}_{\mathbf{G}}\mathbf{P}^{\perp}_{\mathbf{X}_0}$. They are used below in conjunction with Lemma 2.1 to get the following results for $\mathbf{y} \in \mathcal{M}_{\gamma_0}$:

$$\begin{aligned}
\lambda^{(1)}_{\gamma_0} &= \|\mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{X}_0\theta_0\|^2/\tau_0 = 0,\\
\lambda^{(2)}_{\gamma_0} &= \|\mathbf{P}_{\mathbf{X}_0}\mathbf{X}_0\theta_0\|^2/\tau_0 = \|\mathbf{X}_0\theta_0\|^2/\tau_0,\\
\lambda^{(1)}_{\gamma_1} &= \|\mathbf{P}^{\perp}_{\mathbf{X}}\mathbf{X}_0\theta_0\|^2/\tau_0 = \|\mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{P}^{\perp}_{\mathbf{G}}\mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{X}_0\theta_0\|^2/\tau_0 = 0,\\
\lambda^{(2)}_{\gamma_1} &= \|\mathbf{P}_{\mathbf{X}}\mathbf{X}_0\theta_0\|^2/\tau_0 = \|(\mathbf{P}_{\mathbf{X}_0} + \mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{P}_{\mathbf{G}}\mathbf{P}^{\perp}_{\mathbf{X}_0})\mathbf{X}_0\theta_0\|^2/\tau_0\\
&= \|\mathbf{X}_0\theta_0\|^2/\tau_0.
\end{aligned}$$

For $\mathbf{y} \in \mathcal{M}_{\gamma_1}$, we have:

$$\begin{aligned}
\lambda^{(1)}_{\gamma_0} &= \|\mathbf{P}^{\perp}_{\mathbf{X}_0}(\mathbf{X}_0\theta_0 + \mathbf{X}_i\theta_i)\|^2/\tau_1 = \|\mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{X}_i\theta_i\|^2/\tau_1,\\
\lambda^{(2)}_{\gamma_0} &= \|\mathbf{P}_{\mathbf{X}_0}(\mathbf{X}_0\theta_0 + \mathbf{X}_i\theta_i)\|^2/\tau_1 = \|\mathbf{X}_0\theta_0 + \mathbf{P}_{\mathbf{X}_0}\mathbf{X}_i\theta_i\|^2/\tau_1,\\
\lambda^{(1)}_{\gamma_1} &= \|\mathbf{P}^{\perp}_{\mathbf{X}}\mathbf{X}[\theta_0^{\top}\ \theta_i^{\top}]^{\top}\|^2/\tau_1 = 0,\\
\lambda^{(2)}_{\gamma_1} &= \|\mathbf{P}_{\mathbf{X}}\mathbf{X}[\theta_0^{\top}\ \theta_i^{\top}]^{\top}\|^2/\tau_1 = \|\mathbf{X}_0\theta_0 + \mathbf{X}_i\theta_i\|^2/\tau_1.
\end{aligned}$$

ii) When $\mathbf{y} \in \mathcal{M}_{\gamma_0}$ and $n > m+2$, we apply the results above and the formula (30.3a) from [5], to calculate:

$$\begin{aligned}
&E\left[F''_{|\gamma_0|,n-|\gamma_0|}(\lambda^{(2)}_{\gamma_0},\lambda^{(1)}_{\gamma_0})\right]\\
&\quad= E\left[F'_{m_0,n-m_0}(\lambda^{(2)}_{\gamma_0})\right]\\
&\quad= \frac{(n-m_0)(m_0+\lambda^{(2)}_{\gamma_0})}{m_0(n-m_0-2)}\\
&\quad= \left[1+\frac{2}{n-m_0-2}\right]\left[1+\frac{\|\mathbf{X}_0\theta_0\|^2/\tau_0}{m_0}\right],\\
&E\left[F''_{|\gamma_1|,n-|\gamma_1|}(\lambda^{(2)}_{\gamma_1},\lambda^{(1)}_{\gamma_1})\right]\\
&\quad= E\left[F'_{m,n-m}(\lambda^{(2)}_{\gamma_1})\right]\\
&\quad= \left[1+\frac{2}{n-m-2}\right]\left[1+\frac{\|\mathbf{X}_0\theta_0\|^2/\tau_0}{m}\right].
\end{aligned}$$

Similarly, for $\mathbf{y} \in \mathcal{M}_{\gamma_1}$ and $n > m+2$, we have:

$$\begin{aligned}
&E\left[F''_{|\gamma_0|,n-|\gamma_0|}(\lambda^{(2)}_{\gamma_0},\lambda^{(1)}_{\gamma_0})\right]\\
&\quad\approx \frac{1}{1+\lambda^{(1)}_{\gamma_0}/(n-m_0)}E\left[F'_{m_0,n-m_0}(\lambda^{(2)}_{\gamma_0})\right] \qquad (24)\\
&\quad= \left[1+\frac{2}{n-m_0-2}\right]\frac{1+\frac{\|\mathbf{X}_0\theta_0+\mathbf{P}_{\mathbf{X}_0}\mathbf{X}_i\theta_i\|^2/\tau_1}{m_0}}{1+\frac{\|\mathbf{P}^{\perp}_{\mathbf{X}_0}\mathbf{X}_i\theta_i\|^2/\tau_1}{n-m_0}},\\
&E\left[F''_{|\gamma_1|,n-|\gamma_1|}(\lambda^{(2)}_{\gamma_1},\lambda^{(1)}_{\gamma_1})\right]\\
&\quad= E\left[F'_{m,n-m}(\lambda^{(2)}_{\gamma_1})\right]\\
&\quad= \left[1+\frac{2}{n-m-2}\right]\left[1+\frac{\|\mathbf{X}_0\theta_0+\mathbf{X}_i\theta_i\|^2/\tau_1}{m}\right].
\end{aligned}$$

The approximation (24) is taken from [5], where it was derived under the assumption that $\frac{\lambda^{(1)}_{\gamma_0}}{n-|\gamma_0|} < \frac{1}{2}$, which in our case reduces to the condition (20).

□

REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, AC-19:716–723, Dec. 1974.

[2] P. Djuric. Asymptotic MAP criteria for model selection. *IEEE Trans. Signal. Proces.*, 46(10):2726–2735, Oct. 1998.

[3] F. Graybill. *An introduction to linear statistical models*. McGraw-Hill Book Company, 1961.

[4] M. Hansen and B. Yu. Minimum Description Length model selection criteria for generalized linear models. In D. Goldstein, editor, *Science and statistics: a festchrift for Terry Speed*, volume 40, pages 145–164. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 2002.

[5] N. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions*, volume 2. John Wiley & Sons, second edition, 1995.

[6] S. Kay. *Fundamentals of statistical signal processing: detection theory*. Prentice Hall, 1998.

[7] S. Kay. Conditional model order estimation. *IEEE Trans. Signal. Proces.*, 49(9):1910–1917, Sep. 2001.

[8] S. Kay. Exponentially embedded families-new approaches to model order estimation. *IEEE Trans. on Aerospace and Electronic Systems*, 41(1):333–345, Jan. 2005.

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

[10] G. Qian and H. Künsch. Some notes on Rissanen's stochastic complexity. *IEEE Trans. Inf. Theory*, 44(2):782–786, Mar. 1998.

[11] S. Razavi and C. Giurcăneanu. Optimally distinguishable distributions: a new approach to composite hypothesis testing with applications to the classical linear model. *IEEE Trans. Signal. Proces. (to appear)*, doi: 10.1109/TSP.2009.2017568.

[12] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[13] J. Rissanen. Hypothesis selection and testing by the MDL principle. *Computer Journal*, 42(4):260–269, 1999.

[14] J. Rissanen. MDL denoising. *IEEE Trans. Inf. Theory*, 46(7):2537–2543, Nov. 2000.

[15] J. Rissanen. *Information and Complexity in Statistical Modeling*. Springer Verlag, 2007.

[16] L. Scharf and B. Friedlander. Matched subspace detectors. *IEEE Trans. Signal. Proces.*, 42(8):2146–2157, 1994.

[17] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, Mar. 1978.

[18] G. Seber and A. Lee. *Linear regression analysis*. Wiley-Interscience, 2003.

[19] T. Söderström. On model structure testing in system identification. *Int. J. Control*, 26(1):1–18, 1977.

[20] P. Stoica and Y. Selen. A review of information criterion rules. *IEEE Signal. Proces. Mag.*, 21(4):36–47, Jul. 2004.

[21] P. Stoica, Y. Selen, and J. Li. On information criteria and the generalized likelihood ratio test of model order selection. *IEEE Signal Processing Letters*, 11(10):794–797, 2004.