Signal Processing 91 (2011) 1671–1692

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Review

# Variable selection in linear regression: Several approaches based on normalized maximum likelihood

Ciprian Doru Giurcăneanu *, Seyed Alireza Razavi, Antti Liski

*Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland*

ARTICLE INFO

*Article history:*
Received 2 June 2010
Received in revised form
14 March 2011
Accepted 21 March 2011
Available online 1 April 2011

*Keywords:*
Gaussian linear regression
Model selection
Normalized maximum likelihood
Rhomboidal constraint
Ellipsoidal constraint

ABSTRACT

The use of the normalized maximum likelihood (NML) for model selection in Gaussian linear regression poses troubles because the normalization coefficient is not finite. The most elegant solution has been proposed by Rissanen and consists in applying a particular constraint for the data space. In this paper, we demonstrate that the methodology can be generalized, and we discuss two particular cases, namely the rhomboidal and the ellipsoidal constraints. The new findings are used to derive four NML-based criteria. For three of them which have been already introduced in the previous literature, we provide a rigorous analysis. We also compare them against five state-of-the-art selection rules by conducting Monte Carlo simulations for families of models commonly used in signal processing. Additionally, for the eight criteria which are tested, we report results on their predictive capabilities for real life data sets.

**Contents**

* Corresponding author. Tel.: +358 3 3115 3832; fax: +358 3 3115 4989.
*E-mail addresses:* ciprian.giurcaneanu@tut.fi (C.D. Giurcăneanu), alireza.razavi@tut.fi (S.A. Razavi), antti.liski@tut.fi (A. Liski).

0165-1684/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2011.03.015

## 1. Introductory remarks and problem formulation

One of the fundamental research topics addressed in signal processing is the linear least-squares regression problem. Let the measurements $\mathbf{y} \in \mathbb{R}^{n\times 1}$ be modeled by

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

where $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the regressor matrix having more rows than columns ($n > m$), $\boldsymbol{\beta} \in \mathbb{R}^{m\times 1}$ is the vector of unknown parameters, and the entries of $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times 1}$ are samples from an independent and identically distributed (i.i.d.) Gaussian process of zero-mean and variance $\tau$. Hereafter, 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.

Because in most of the practical applications, not all the parameters $\beta_1,\ldots,\beta_m$ are equally important in modeling $\mathbf{y}$, one wants to eliminate those that are deemed to be irrelevant. This reduces to choose a subset of the regressor variables indexed by $\gamma \subseteq \{1,\ldots,m\}$. It is customary to select $\gamma$ by using either the Akaike Information Criterion (AIC) [1], or the Bayesian Information Criterion (BIC) [29]. Both AIC and BIC can be seen like particular cases of a more general class of asymptotic criteria which are expressed as the sum of two terms: the first one 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 [36, Appendix C].

It is widely recognized that BIC is equivalent with an information theoretic criterion called MDL (minimum description length) [23]. However, MDL is not only a simple formula, but it is a principle [8].

To show how the most recent MDL-based developments can be applied to the linear regression problem, we focus on the computation of the stochastic complexity (SC) [25,26]. Let $\boldsymbol{\beta}_\gamma \in \mathbb{R}^{k\times 1}$ be the vector of the unknown regression coefficients within the $\gamma$-subset. We denote the cardinality of $\gamma$ by $k$, and we make the assumption that $k$ is strictly positive. The case $k=0$ will be treated separately. The matrix $\mathbf{X}_\gamma$ is given by the columns of $\mathbf{X}$ that correspond to the $\gamma$-subset. Similarly with (1), we have

$$\mathbf{y} = \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma + \boldsymbol{\varepsilon}_\gamma, \tag{2}$$

where the entries of $\boldsymbol{\varepsilon}_\gamma$ are Gaussian distributed with zero-mean and unknown variance $\tau_\gamma$. Under the hypothesis that $\mathbf{X}_\gamma$ has full-rank, the maximum likelihood (ML) estimates are [31]: $\hat{\boldsymbol{\beta}}_\gamma(\mathbf{y}) = (\mathbf{X}_\gamma^\top\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^\top\mathbf{y}$ and $\hat{\tau}_\gamma(\mathbf{y}) = \|\mathbf{y}-\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2/n$, where the superscripts $(\cdot)^\top$ and $(\cdot)^{-1}$ denote the transpose and the matrix inverse, respectively. The operator $\|\cdot\|$ is employed for the Euclidean norm. Whenever it is clear from the context which measurements are used for estimation, the simpler notation $\hat{\boldsymbol{\beta}}_\gamma$ will be preferred to $\hat{\boldsymbol{\beta}}_\gamma(\mathbf{y})$. The same applies for the use of $\hat{\tau}_\gamma$ instead of $\hat{\tau}_\gamma(\mathbf{y})$. To evaluate the SC for the data vector $\mathbf{y}$, given the $\gamma$-structure, we have to compute

$$\mathrm{SC}(\mathbf{y};\gamma) = L(\mathbf{y};\gamma) + L(a,b) + \frac{n}{2}\ln(n\pi), \tag{3}$$

$$L(\mathbf{y};\gamma) = \frac{n-k}{2}\ln\hat{\tau}_\gamma + \frac{k}{2}\ln\frac{\|\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2}{n} - \ln\Gamma\left(\frac{n-k}{2}\right) - \ln\Gamma\left(\frac{k}{2}\right), \tag{4}$$

$$L(a,b) = 2\ln\ln\frac{b}{a}, \tag{5}$$

where $\ln(\cdot)$ denotes the natural logarithm and $\Gamma(\cdot)$ is the Euler integral of the second kind. Additionally, the real-valued hyper-parameters $a$ and $b$ satisfy the condition: $b > a$.

The complete formula includes also the description length for the $\gamma$-structure, $L(\gamma)$, whose expression is given in [26]. Because in many practical problems, the term $L(\gamma)$ has a marginal effect, we will ignore it. Example 4 in Section 4 will be the only case when we will consider the contribution of this term. For clarifications on the role of $L(\gamma)$, see [27].

The case $k=0$ is equivalent to $\gamma = \emptyset$, and occurs when the observations $\mathbf{y}$ are assumed to be pure Gaussian noise with zero-mean and unknown variance. In this situation, the stochastic complexity takes the particular form

$$\mathrm{SC}(\mathbf{y};\emptyset) = L(\mathbf{y};\emptyset) + \frac{1}{2}L(a,b) + \frac{n}{2}\ln(n\pi), \tag{6}$$

$$L(\mathbf{y};\emptyset) = \frac{n}{2}\ln\frac{\|\mathbf{y}\|^2}{n} - \ln\Gamma\left(\frac{n}{2}\right), \tag{7}$$

where $L(a,b)$ is defined in (5). In this work, we neglect the terms given by $L(a,b)$. We refer to [7,26, Section 9.3] for a more elaborated discussion on the conditions when $2\ln\ln(b/a)$ and $\ln\ln(b/a)$ can be dropped from (3) and (6), respectively.

In line with the MDL principle, selection of the best structure amounts to evaluate $\mathrm{SC}(\mathbf{y};\gamma)$ for all $\gamma \subseteq \{1,\ldots,m\}$, and then to pick-up the subset that minimizes the stochastic complexity. Another information theoretic criterion which is akin to formulas in (3)–(5) and (6)–(7) has been derived in [9,10] by using a universal mixture model. More interestingly, Kay has proposed in [16] a selection rule based on exponentially embedded families (EEF) of probability density functions, and which is similar to the one introduced by Rissanen in [25]. A comparison of the criteria from [10,16,25], for the case when the noise variance is assumed to be known, can be found in [6, Section 3.3]. The minimum message length (MML) principle was recently used in [28] to yield two new model selection criteria, and it turned out that both of them are closely related to SC.

The fact that formulas which are almost the same with the expression of SC can be obtained by various approaches is indeed an indicator for the practitioner that the use of SC might be the right choice. However, for the work presented in this paper, the most important is not the SC formula, but the methodology applied by Rissanen for its derivation. The central role is played by the normalized maximum likelihood (NML) density function:

$$\hat{f}(\mathbf{y};\gamma)=\frac{f(\mathbf{y};\hat{\boldsymbol{\beta}}_{\gamma}(\mathbf{y}),\hat{\tau}_{\gamma}(\mathbf{y}))}{C(\gamma)}, \tag{8}$$

$$C(\gamma)=\int f(\mathbf{y};\hat{\boldsymbol{\beta}}_{\gamma}(\mathbf{y}),\hat{\tau}_{\gamma}(\mathbf{y}))\,\mathrm{d}\mathbf{y}, \tag{9}$$

where $f(\mathbf{y};\hat{\boldsymbol{\beta}}_{\gamma}(\mathbf{y}),\hat{\tau}_{\gamma}(\mathbf{y}))$ is the ML. In the equation above, the domain of integration is the entire space of observations. Note also that (9) gives the definition of the *parametric complexity*. It was shown in [25,26] that NML has two important optimality properties which recommend it to be used in the evaluation of SC. More precisely, SC is computed as the code length associated with NML: $\mathrm{SC}(\mathbf{y};\gamma)=-\ln\hat{f}(\mathbf{y};\gamma)$. The key point is that the parametric complexity in the linear regression case is not finite, or equivalently, the integral in (9) is not finite. To circumvent this difficulty, Rissanen proposed in [25] to constrain the integration domain in the space of observations such that the integral becomes finite, and this led to the criterion given by the formulas in (3)–(5) and (6)–(7).

We note in passing that, according to Scopus, Ref. [25] has been cited more than 50 times. Hence, the SC-criterion is widely used, and one of the reasons is the following. The criterion is independent of arbitrarily selected hyper-parameters if the terms that involve $L(a,b)$ are neglected. Surprisingly, for about one decade, it was totally ignored the important fact that the closed-form expression of the criterion depends on the particular constraint which has been involved in its derivation. Only recently, it was shown in [18] that two other criteria can be obtained by employing constraints which are different of the one used in [25].

The most recent findings lead to the conclusion that novel NML-based criteria can be devised by enforcing various constraints. However, in the previous literature, it was not investigated how the selection of the constraint influences the performance of the resulting criterion. To fill the gap, this paper provides the following results:

(i) We demonstrate in Section 2 that the methodology introduced by Rissanen can be applied in a more general framework, and not only for the ellipsoidal constraints which have been considered in [18,25]. In the same section, we study the particular case of rhomboidal constraint.
(ii) In Section 3, we conduct a rigorous analysis of the relationship between Rissanen criterion and the two criteria that have been introduced in [18].
(iii) Section 4 is devoted to numerical examples which compare the capabilities of the NML-based selection rules against other criteria. The experiments are performed with simulated data as well as real life data sets.

Conclusions are outlined in Section 5, where we also give some guidance on the use of various criteria in model selection.

## 2. Parametric complexity with constraints

### 2.1. General case

To simplify the notations, we drop the index $\gamma$ when discussing the general case. Let us define $\mathcal{Y}_{\rho}(R,\tau_0)=\{\mathbf{y}:\rho(\hat{\boldsymbol{\beta}})\le R,\hat{\tau}\ge\tau_0\}$, where $R$ and $\tau_0$ are strictly positive. The mapping $\rho:\mathbb{R}^k\to\mathbb{R}$ is chosen such that, for all $R>0$, the set $\mathcal{B}_{\rho}(R)=\{\hat{\boldsymbol{\beta}}:\rho(\hat{\boldsymbol{\beta}})\le R\}$ is convex and its volume $V_{\rho}(R)=\int_{\mathcal{B}_{\rho}(R)}\mathrm{d}\hat{\boldsymbol{\beta}}$ has the expression

$$V_{\rho}(R)=\eta R^{\zeta k}. \tag{10}$$

The constant $\eta$ is strictly positive and, in some cases, it might depend on the regressor matrix $\mathbf{X}$. Additionally, the constant $\zeta$ is also assumed to be strictly positive.

Hence, the definition of NML from (8)–(9) is transformed to

$$\hat{f}_{\rho}(\mathbf{y};R,\tau_0)=\frac{f(\mathbf{y};\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y}))}{C_{\rho}(R,\tau_0)}, \tag{11}$$

$$C_{\rho}(R,\tau_0)=\int_{\mathcal{Y}_{\rho}(R,\tau_0)} f(\mathbf{y};\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y}))\,\mathrm{d}\mathbf{y}. \tag{12}$$

It is well known that the numerator in (11) is given by [31]

$$f(\mathbf{y};\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y}))=[2\pi\hat{\tau}\exp(1)]^{-n/2}. \tag{13}$$

For the denominator, we prove in Appendix A that

$$C_{\rho}(R,\tau_0)=(2A_{n,k}/k)\tau_0^{-k/2}\eta R^{\zeta k}, \tag{14}$$

where

$$A_{n,k}=\frac{|\mathbf{X}^{\top}\mathbf{X}|^{1/2}}{(n\pi)^{k/2}}\frac{\left(\frac{n}{2\exp(1)}\right)^{n/2}}{\Gamma\left(\frac{n-k}{2}\right)}. \tag{15}$$

The operator $|\cdot|$ denotes the determinant of the matrix in the argument.

Remark in (14) that the normalizing constant $C_{\rho}(R,\tau_0)$ becomes smaller when $R$ decreases. Because we want to minimize the code length given by $-\ln\hat{f}_{\rho}(\mathbf{y};R,\tau_0)=-\ln f(\mathbf{y};\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y}))+\ln C_{\rho}(R,\tau_0)$, we assign to $R$ the smallest possible value, namely $R=\tilde{R}$, where $\tilde{R}=\rho(\hat{\boldsymbol{\beta}})$. We choose $\tilde{\tau}_0=\hat{\tau}$ like in [26], and the expression from (11) becomes

$$\hat{f}_{\rho}(\mathbf{y};\tilde{R},\tilde{\tau}_0)=\frac{f(\mathbf{y};\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y}))}{C_{\rho}(\tilde{R},\tilde{\tau}_0)}. \tag{16}$$

Then we perform the second normalization step. Let $\mathcal{Y}(R_1,R_2,\tau_1,\tau_2)=\{\mathbf{y}:R_1\le\rho(\hat{\boldsymbol{\beta}}(\mathbf{y}))\le R_2,\tau_1\le\hat{\tau}(\mathbf{y})\le\tau_2\}$, where $R_2>R_1>0$ and $\tau_2>\tau_1>0$. By using (16), we have

$$\hat{f}_{\rho}(\mathbf{y})=\frac{\hat{f}_{\rho}(\mathbf{y};\tilde{R},\tilde{\tau}_0)}{\overline{C}_{\rho}(R_1,R_2,\tau_1,\tau_2)}=\frac{f(\mathbf{y};\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y}))/C_{\rho}(\tilde{R},\tilde{\tau}_0)}{\overline{C}_{\rho}(R_1,R_2,\tau_1,\tau_2)}. \tag{17}$$

The normalizing constant is given by

$$\overline{C}_{\rho}(R_1,R_2,\tau_1,\tau_2)=\int_{\mathcal{Y}(R_1,R_2,\tau_1,\tau_2)}\hat{f}_{\rho}(\mathbf{y};\tilde{R},\tilde{\tau}_0)\,\mathrm{d}\mathbf{y}, \tag{18}$$

and after some calculations which are outlined in Appendix A, we obtain

$$\overline{C}_\rho(R_1,R_2,\tau_1,\tau_2)=\frac{\zeta k^2}{2}\ln\frac{\tau_2}{\tau_1}\ln\frac{R_2}{R_1}. \tag{19}$$

We collect the results from (13), (14), (17) and (19) to get the expression of the negative logarithm of NML, when the mapping $\rho(\cdot)$ is used to define the constraint for the evaluation of the parametric complexity:

$$\begin{aligned}-\ln\hat{f}_\rho(\mathbf{y})&=-\ln f(\mathbf{y};\hat{\boldsymbol{\beta}},\hat{\tau})+\ln C_\rho(\rho(\hat{\boldsymbol{\beta}}),\hat{\tau})+\ln\overline{C}_\rho(R_1,R_2,\tau_1,\tau_2)\\&=\frac{n-k}{2}\ln\hat{\tau}+\zeta k\ln\rho(\hat{\boldsymbol{\beta}})-\ln\Gamma\left(\frac{n-k}{2}\right)+\ln\left[\zeta k\frac{\eta|\mathbf{X}^\top\mathbf{X}|^{1/2}}{(n\pi)^{k/2}}\right]\\&\quad+\frac{n}{2}\ln(n\pi)+\ln\left(\ln\frac{\tau_2}{\tau_1}\ln\frac{R_2}{R_1}\right).\end{aligned}\tag{20}$$

It is obvious that, in the equations above, we have $\hat{\tau}=\hat{\tau}_\gamma$, $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}_\gamma$ and $\mathbf{X}=\mathbf{X}_\gamma$. Conventionally we take $\tau_1=R_1=a$ and $\tau_2=R_2=b$, where $b>a>0$. So,

$$-\ln\hat{f}_\rho(\mathbf{y})=L_\rho(\mathbf{y};\gamma)+L(a,b)+\frac{n}{2}\ln(n\pi), \tag{21}$$

$$L_\rho(\mathbf{y};\gamma)=\frac{n-k}{2}\ln\hat{\tau}_\gamma+\zeta k\ln\rho(\hat{\boldsymbol{\beta}}_\gamma)-\ln\Gamma\left(\frac{n-k}{2}\right)+\ln\left[\zeta k\frac{\eta|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|^{1/2}}{(n\pi)^{k/2}}\right], \tag{22}$$

where $L(a,b)$ is the same as in (5).

For the sake of completeness, we consider also an approximate formula for the negative logarithm of NML [24]:

$$-\ln\hat{f}(\mathbf{y})=-\ln f(\mathbf{y};\hat{\boldsymbol{\beta}},\hat{\tau})+\frac{k+1}{2}\ln\frac{n}{2\pi}+\ln\int|\mathbf{J}_\infty(\boldsymbol{\beta},\tau)|^{1/2}\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\tau+o(1), \tag{23}$$

where

$$\mathbf{J}_\infty(\boldsymbol{\beta},\tau)=\lim_{n\to\infty}\mathbf{J}_n(\boldsymbol{\beta},\tau), \tag{24}$$

$$\mathbf{J}_n(\boldsymbol{\beta},\tau)=\begin{bmatrix}(\mathbf{X}^\top\mathbf{X})/(n\tau) & \mathbf{0}\\ \mathbf{0} & 1/(2\tau^2)\end{bmatrix}. \tag{25}$$

Remark in (23)–(25) that we have dropped the index $\gamma$. In (25), we have used the expression (see, for example, [14]) of the Fisher information matrix (FIM) for the linear model in (2). Note that, for many models used in signal processing, the right-hand side of (24) has a finite limit [36, Appendix C]. On contrary, the value of the integral in (23) is not finite if the domain of integration is the entire parameter space. This problem is well known and some of the proposed solutions involve arbitrarily chosen restrictions for the ranges of the parameters. A comprehensive discussion on this issue can be found in [11]. We demonstrate in Appendix A how the difficulty can be circumvented by applying constraints similar with those employed to get (20).

### 2.2. Rissanen formula

The constraint used by Rissanen is $\rho_1(\hat{\boldsymbol{\beta}}_\gamma)\le R$, where $\rho_1(\hat{\boldsymbol{\beta}}_\gamma)=\|\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2/n$ [26]. This makes the volume $V_{\rho_1}(R)$ to be given by (10) with $\eta=(n\pi)^{k/2}/[(k/2)\Gamma(k/2)|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|^{1/2}]$ and $\zeta=1/2$. It is a simple exercise to show that, for the particular case when $\rho(\hat{\boldsymbol{\beta}}_\gamma)=\rho_1(\hat{\boldsymbol{\beta}}_\gamma)$, the formula in (21)–(22) is identical with the one from (3)–(5). The expression of SC can be further simplified by operating the following modifications: (i) neglect the constant term $(n/2)\ln(n\pi)$ and the term $L(a,b)$; (ii) use the Stirling approximation (see Appendix A and [26,27])

$$\ln\Gamma(z)=(z-\tfrac{1}{2})\ln z-z+\tfrac{1}{2}\ln(2\pi), \tag{26}$$

and then discard all terms which do not depend on the $\gamma$-structure; (iii) multiply by two the resulting criterion. This leads to

$$\mathrm{SC}_{\rho_1}(\mathbf{y};\gamma)=(n-k)\ln\frac{\hat{\tau}_\gamma}{n-k}+k\ln\frac{\|\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2/n}{k}+\ln[k(n-k)]. \tag{27}$$

The above form of SC is the one which appeared most frequently in the literature after it was introduced in [25].

### 2.3. Rhomboidal constraint

Consider the constraint $\rho_0(\hat{\boldsymbol{\beta}}_\gamma)\le R$, where $\rho_0(\hat{\boldsymbol{\beta}}_\gamma)$ is given by the 1-norm of $\hat{\boldsymbol{\beta}}_\gamma$, and we write $\rho_0(\hat{\boldsymbol{\beta}}_\gamma)=\|\hat{\boldsymbol{\beta}}_\gamma\|_1$. The region defined by the constraint is a diamond when $k=2$, and it becomes a rhomboid when $k>2$ [12]. The volume $V_{\rho_0}(R)$ can be computed by observing that

$$V_{\rho_0}(R)=2^k\times\int_{\substack{\hat{\beta}_1,\ldots,\hat{\beta}_k\ge 0\\ \hat{\beta}_1+\cdots+\hat{\beta}_k\le R}}\mathrm{d}\hat{\boldsymbol{\beta}}$$

because of the symmetry. Then we get $V_{\rho_0}(R)=(2R)^k/k!$. The result is easily verified for $k\in\{1,2\}$ and is proven for any $k>2$ by mathematical induction. More importantly, the formula which gives the volume $V_{\rho_0}(R)$ can be obtained from the one in (10) by choosing $\eta=2^k/k!$ and $\zeta=1$. Hence, we can get a new NML-based criterion by using in (21)–(22) the definition of $\rho_0(\cdot)$. For writing more compactly the new selection rule, we multiply by two the expression in (21), and we ignore the sum $2L(a,b)+n\ln(n\pi)$. Some elementary calculations lead to

$$\begin{aligned}\mathrm{SC}_{\rho_0}(\mathbf{y};\gamma)&=(n-k)\ln\hat{\tau}_\gamma+k\ln\frac{\|\hat{\boldsymbol{\beta}}_\gamma\|_1^2}{n}-2\ln\Gamma\left(\frac{n-k}{2}\right)-2\ln\Gamma(k)\\&\quad+\ln\frac{4^k|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|}{\pi^k}.\end{aligned}$$

We modify the formula above by applying the Stirling approximation from (26), and by discarding the sum $(n-2)\ln 2+n-2\ln(2\pi)$, which was also neglected in (27). Thus, we have

$$\begin{aligned}\mathrm{SC}_{\rho_0}(\mathbf{y};\gamma)&=(n-k)\ln\frac{\hat{\tau}_\gamma}{n-k}+k\ln\frac{\|\hat{\boldsymbol{\beta}}_\gamma\|_1^2/n}{k}+\ln[k(n-k)]\\&\quad+k\ln\frac{2\exp(1)}{\pi k}+\ln(2|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|).\end{aligned}\tag{28}$$

From (27) and (28), it is obvious that the goodness-of-fit term is the same for both $\mathrm{SC}_{\rho_1}(\mathbf{y};\gamma)$ and $\mathrm{SC}_{\rho_0}(\mathbf{y};\gamma)$. We want to check which is the relationship between the penalty terms of the two criteria. For ease of comparison, we assume that the columns of $\mathbf{X}_\gamma$ are the first $k$ columns of the $n\times n$ identity matrix, which implies

$$\mathrm{PEN}_{\rho_0}(\mathbf{y};\gamma)-\mathrm{PEN}_{\rho_1}(\mathbf{y};\gamma)=k\ln\frac{2\exp(1)\cos^2(\alpha_\gamma)}{\pi}+\ln 2, \tag{29}$$

where $\cos\alpha_\gamma = \|\hat{\boldsymbol{\beta}}_\gamma\|_1/(\sqrt{k}\|\hat{\boldsymbol{\beta}}_\gamma\|)$. Equivalently, $\alpha_\gamma$ is the angle between the vector $[|\hat{\beta}_1|,\ldots,|\hat{\beta}_k|]^\top$, which is given by the magnitudes of the estimates, and the vector $[1,\ldots,1]^\top$. Remark in (29) that $\mathrm{PEN}_{\rho_0}(\mathbf{y};\gamma)-\mathrm{PEN}_{\rho_1}(\mathbf{y};\gamma)>0$ if and only if $\alpha_\gamma<\arccos(\mathrm{Th}_k)$, where $\mathrm{Th}_k=\{[\pi/(2\exp(1))](1/2^{1/k})\}^{1/2}$. For all $k\geq 1$, the inequality $\mathrm{Th}_k\in(0,1)$ is satisfied, and for $k\gg 1$, we have $\arccos(\mathrm{Th}_k)\approx(2\pi)/9$.

To gain more insight, we assume that $\mathbf{y}\sim\mathcal{N}(\overline{\boldsymbol{\beta}},\overline{\tau}\mathbf{I})$, where $\mathcal{N}(\boldsymbol{\mu},\mathbf{R})$ denotes the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{R}$. The vector $\overline{\boldsymbol{\beta}}$ is chosen such that to have $\overline{k}$ entries equal to a non-zero constant $\beta$, and all other entries are zeros. Additionally, $\overline{k}\ll n$, and the value of $\overline{\tau}$ is selected to guarantee a certain signal-to-noise ratio. Let $y_{(1)},\ldots,y_{(n)}$ be the measurements sorted in the decreasing order of their magnitudes. For each $k\in\{1,\ldots,n-2\}$, we define the structure $\gamma_k=\{(1),\ldots,(k)\}$ such that $\hat{\boldsymbol{\beta}}_{\gamma_k}=[y_{(1)},\ldots,y_{(k)}]^\top$ and $\hat{\tau}_{\gamma_k}=(1/n)\sum_{i=k+1}^{n}y_{(i)}^2$. When $k>\overline{k}$, if $k$ increases, then the angle $\alpha_{\gamma_k}$ increases also, and $\mathrm{PEN}_{\rho_0}(\mathbf{y};\gamma_k)$ becomes smaller than $\mathrm{PEN}_{\rho_1}(\mathbf{y};\gamma_k)$. Hence, the criterion $\mathrm{SC}_{\rho_0}$ penalizes less than $\mathrm{SC}_{\rho_1}$ when $k$ is large, which makes to be more likely that $\mathrm{SC}_{\rho_1}$ selects a sparser solution, and not $\mathrm{SC}_{\rho_0}$.

This outcome is surprising because it is known from the previous literature [12, Chapter 3] that the selection rules which have as penalty term the 1-norm of the vector of estimates are prone to pick-up the sparse solutions.

The formulas derived with the general methodology described in Section 2.1 must be used with caution in practice, and only after their properties are carefully investigated. Next, we focus on two other NML-based criteria, which have been introduced in [18] to cope with the presence of collinearity.

## 3. Ellipsoidal constraint

### 3.1. Formulas from [18]

The solution proposed in [18] for the computation of the parametric complexity relies on the following ellipsoidal constraint: $(\hat{\boldsymbol{\beta}}_\gamma^\top\mathbf{Q}\hat{\boldsymbol{\beta}}_\gamma)/n\leq R$, where the matrix $\mathbf{Q}$ is chosen to be symmetric and positive definite. By applying the formula for the volume of an ellipsoid [30], it is easy to verify for $\rho(\hat{\boldsymbol{\beta}}_\gamma)=\hat{\boldsymbol{\beta}}_\gamma^\top\mathbf{Q}\hat{\boldsymbol{\beta}}_\gamma$ that $V_\rho(R)$ is a particular case of (10) for which $\eta=(n\pi)^{k/2}/[(k/2)\Gamma(k/2)|\mathbf{Q}|^{1/2}]$ and $\zeta=1/2$. By employing in (21)–(22) the expressions of $\eta$ and $\zeta$, we have

$$-\ln\hat{f}_\rho(\mathbf{y})=\frac{n-k}{2}\ln\hat{\tau}_\gamma+\frac{k}{2}\ln\frac{\hat{\boldsymbol{\beta}}_\gamma^\top\mathbf{Q}\hat{\boldsymbol{\beta}}_\gamma}{n}-\ln\Gamma\left(\frac{n-k}{2}\right)$$
$$-\ln\Gamma\left(\frac{k}{2}\right)+\frac{1}{2}\ln\frac{|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|}{|\mathbf{Q}|},\qquad(30)$$

which coincides with [18, Eq. (16)]. Remark in (30) that we have neglected the terms $L(a,b)$ and $(n/2)\ln(n\pi)$.

When $\mathbf{Q}=\mathbf{X}_\gamma^\top\mathbf{X}_\gamma$, the ellipsoidal constraint $(\hat{\boldsymbol{\beta}}_\gamma^\top\mathbf{Q}\hat{\boldsymbol{\beta}}_\gamma)/n\leq R$ is identical with the constraint used by Rissanen, namely $\rho_1(\hat{\boldsymbol{\beta}}_\gamma)\leq R$. Other two possible ways of selecting the matrix $\mathbf{Q}$ have been considered in [18]: $\mathbf{Q}=\mathbf{I}$ and $\mathbf{Q}=(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma)^2$. In the case when $\mathbf{Q}=\mathbf{I}$, the ellipsoidal constraint becomes $\rho_2(\hat{\boldsymbol{\beta}}_\gamma)\leq R$, where $\rho_2(\hat{\boldsymbol{\beta}}_\gamma)=\|\hat{\boldsymbol{\beta}}_\gamma\|^2/n$. By operating in (30) the same type of modifications which allowed to transform (3)–(5) into (27), we get

$$\mathrm{SC}_{\rho_2}(\mathbf{y};\gamma)=(n-k)\ln\frac{\hat{\tau}_\gamma}{n-k}+k\ln\frac{\|\hat{\boldsymbol{\beta}}_\gamma\|^2/n}{k}+\ln[k(n-k)]+\ln|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|.\qquad(31)$$

Similarly, for $\mathbf{Q}=(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma)^2$, the ellipsoidal constraint takes the form $\rho_3(\hat{\boldsymbol{\beta}}_\gamma)\leq R$ with $\rho_3(\hat{\boldsymbol{\beta}}_\gamma)=[\hat{\boldsymbol{\beta}}_\gamma^\top(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma)^2\hat{\boldsymbol{\beta}}_\gamma]/n$, and the corresponding model selection criterion is

$$\mathrm{SC}_{\rho_3}(\mathbf{y};\gamma)=(n-k)\ln\frac{\hat{\tau}_\gamma}{n-k}+k\ln\frac{[\hat{\boldsymbol{\beta}}_\gamma^\top(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma)^2\hat{\boldsymbol{\beta}}_\gamma]/n}{k}$$
$$+\ln[k(n-k)]-\ln|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|.\qquad(32)$$

After discarding the term $-n\ln n$ from the formulas in (27), (31) and (32), we can re-write them as follows. For $i\in\{1,2,3\}$,

$$\mathrm{SC}_{\rho_i}(\mathbf{y};\gamma)=(n-k)\ln S_\gamma^2+k\ln D_\gamma(\mathbf{y};\mathbf{Q}_i)+\ln\frac{n-k}{k^{k-1}},\qquad(33)$$

$$D_\gamma(\mathbf{y};\mathbf{Q}_i)=\frac{\mathbf{y}^\top\mathbf{X}_\gamma(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma)^{-1}\mathbf{Q}_i(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma^\top\mathbf{y}}{(|\mathbf{Q}_i|/|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|)^{1/k}},\qquad(34)$$

where $S_\gamma^2=(n\hat{\tau}_\gamma)/(n-k)$, $\mathbf{Q}_1=\mathbf{X}_\gamma^\top\mathbf{X}_\gamma$, $\mathbf{Q}_2=\mathbf{I}$ and $\mathbf{Q}_3=(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma)^2$. It is evident that all three selection rules have the same goodness-of-fit term, and only $D_\gamma(\mathbf{y};\mathbf{Q}_i)$ makes their penalty terms to be different.

### 3.2. Penalty terms

For better understanding the relationship between the three criteria, we give the following result.

**Proposition 3.1.**

(a) *The equalities*

$$D_\gamma(\mathbf{y};\mathbf{Q}_1)=D_\gamma(\mathbf{y};\mathbf{Q}_2)=D_\gamma(\mathbf{y};\mathbf{Q}_3)\qquad(35)$$

*hold true for all* $\mathbf{y}\in\mathbb{R}^n\backslash\{\mathbf{0}\}$ *if and only if there exists* $q>0$ *such that*

$$\mathbf{Q}_1=q\mathbf{I}.\qquad(36)$$

(b) *If the condition in* (36) *is not satisfied, then for each pair* $(i,j)$ *with the property that* $1\leq i<j\leq 3$, *the sign of the difference*

$$D_\gamma(\mathbf{y};\mathbf{Q}_i)-D_\gamma(\mathbf{y};\mathbf{Q}_j)$$

*is not the same for all* $\mathbf{y}\in\mathbb{R}^n\backslash\{\mathbf{0}\}$.

(c) *For all* $\mathbf{y}\in\mathbb{R}^n\backslash\{\mathbf{0}\}$, *we have*

$$\max\{D_\gamma(\mathbf{y};\mathbf{Q}_2),D_\gamma(\mathbf{y};\mathbf{Q}_3)\}\geq D_\gamma(\mathbf{y};\mathbf{Q}_1).\qquad(37)$$

Proof is deferred to Appendix B.

From the proposition above, we see that the criteria $\mathrm{SC}_{\rho_1}(\mathbf{y};\gamma)$, $\mathrm{SC}_{\rho_2}(\mathbf{y};\gamma)$ and $\mathrm{SC}_{\rho_3}(\mathbf{y};\gamma)$ are identical only when the columns of the matrix $\mathbf{X}_\gamma$ are orthogonal and the 2-norm is the same for all of them. In general, it is not possible to claim that one criterion has a penalty term which is stronger than the penalty terms of the others. However, the inequality in (37) guarantees that at least one of the criteria $\mathrm{SC}_{\rho_2}(\mathbf{y};\gamma)$ and $\mathrm{SC}_{\rho_3}(\mathbf{y};\gamma)$ has a penalty term which is stronger than the penalty term of the Rissanen criterion.

Next we investigate the behavior of the three selection rules for the case when the matrix $\mathbf{X}_\gamma$ is rank deficient. Let us use the notation $\mathbf{X}_k$ instead of $\mathbf{X}_\gamma$. Furthermore, we partition the matrix into two blocks: $\mathbf{X}_k = [\mathbf{X}_{k\text{-}1}\ \mathbf{x}_k]$. Note that $\mathbf{X}_{k\text{-}1}$ contains the first $k-1$ columns of $\mathbf{X}_k$. We assume that $\mathbf{X}_{k\text{-}1}$ has full-rank, and the source of rank deficiency for $\mathbf{X}_k$ is the fact that the linear subspaces $\langle \mathbf{X}_{k\text{-}1} \rangle$ and $\langle \mathbf{x}_k \rangle$ are "very close" to each other. For a full-rank matrix $\mathbf{M}$ having more rows than columns, $\langle \mathbf{M} \rangle$ is the column space of $\mathbf{M}$.

To measure the "closeness", we employ the *principal angle* $\alpha \in [0,\pi/2]$ between $\langle \mathbf{X}_{k\text{-}1} \rangle$ and $\langle \mathbf{x}_k \rangle$ [3]. If the columns of $\mathbf{U}_{k-1}$ form a unitary basis for $\langle \mathbf{X}_{k\text{-}1} \rangle$ and $\mathbf{u}_k$ is a unitary basis for $\langle \mathbf{x}_k \rangle$, then $\cos\alpha$ is the singular value of $\mathbf{U}_{k-1}^\top \mathbf{u}_k$. Eq. (13) from [3] guarantees that there exists $\mathbf{w} \in \mathbb{R}^{n\times 1}$ with $\|\mathbf{w}\| = 1$ such that

$$\mathbf{P}_{k\text{-}1}^\perp \mathbf{x}_k = \sin(\alpha)\|\mathbf{x}_k\|\mathbf{w}, \tag{38}$$

where $\mathbf{P}_{k\text{-}1}^\perp = \mathbf{I} - \mathbf{P}_{k\text{-}1}$ and $\mathbf{P}_{k\text{-}1} = \mathbf{X}_{k\text{-}1}(\mathbf{X}_{k\text{-}1}^\top \mathbf{X}_{k\text{-}1})^{-1}\mathbf{X}_{k\text{-}1}^\top$ is the orthogonal projection matrix onto the linear subspace $\langle \mathbf{X}_{k\text{-}1} \rangle$. The following proposition clarifies which is the effect of $\alpha \to 0$ on the penalty terms.

**Proposition 3.2.** *If* $\operatorname{rank}(\mathbf{X}_{k\text{-}1}) = k-1$, $\|\mathbf{x}_k\| \neq 0$, $k > 1$ *and* $\mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, *then*:

(a) $\lim_{\alpha \to 0} D_\gamma(\mathbf{y};\mathbf{Q}_1) < \infty$.
(b) $\lim_{\alpha \to 0} D_\gamma(\mathbf{y};\mathbf{Q}_2) = \infty$ *when* $\mathbf{w}^\top \mathbf{y} \neq 0$. *Note that* $\mathbf{w}$ *is defined in* (38).
(c) $\lim_{\alpha \to 0} D_\gamma(\mathbf{y};\mathbf{Q}_3) = \infty$ *when* $\mathbf{X}_k^\top \mathbf{y} \neq \mathbf{0}$.

See Appendix B for the proof.

Remark that, under the assumptions from Proposition 3.2, $\mathrm{SC}_{\rho_2}$ and $\mathrm{SC}_{\rho_3}$ penalize the collinearity more severely than $\mathrm{SC}_{\rho_1}$. The result has to be understood in connection with the fact that variable selection aims to discard those columns of $\mathbf{X}$ which are nearly collinear, and then to use the retained columns for explaining the variation in $\mathbf{y}$ [19, Section 6.7]. This can be nicely formalized by using the *coefficient of determinations* whose definitions are given below.

**Definition 3.1.** Assume that the sum of the entries of $\mathbf{y}$ is zero and $\|\mathbf{y}\| = 1$. Additionally, each column of $\mathbf{X}$ is zero-mean and has unitary Euclidean norm. For an arbitrary $\gamma$-structure with cardinality $k > 0$, we define

$$R^2_{\mathbf{y}\cdot\mathbf{X}_\gamma} = \|\mathbf{P}_\gamma \mathbf{y}\|^2, \tag{39}$$

where $\mathbf{P}_\gamma$ is the orthogonal projection matrix onto the linear subspace $\langle \mathbf{X}_\gamma \rangle$. Moreover, for $i \in \{2,\ldots,k\}$, we have

$$R^2_{i\cdot 1,\ldots,(i-1)} = \|\mathbf{P}_{i-1}\mathbf{x}_i\|^2, \tag{40}$$

where $\mathbf{P}_{i-1}$ denotes the orthogonal projection matrix onto the linear subspace determined by the first $(i-1)$ columns of $\mathbf{X}_\gamma$, and $\mathbf{x}_i$ is the $i$-th column of $\mathbf{X}_\gamma$.

It is clear that (39) and (40) are just particular cases of the general definition that can be found in [19, Section 6.5.2; 31, p. 111]. We emphasize that $R^2_{\mathbf{y}\cdot\mathbf{X}_\gamma}$ is a measure of how much the variance of $\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma$ represents from the total variance of the data $\mathbf{y}$. A similar interpretation can be given for (40).

In the next proposition, we show how the dependence of $\mathbf{y}$ on $\mathbf{X}_\gamma$, as well as the interdependence between the columns of $\mathbf{X}_\gamma$, affect the terms $D_\gamma(\mathbf{y};\mathbf{Q}_1)$, $D_\gamma(\mathbf{y};\mathbf{Q}_2)$ and $D_\gamma(\mathbf{y};\mathbf{Q}_3)$.

**Proposition 3.3.** *When* $\mathbf{X}_\gamma$ *and* $\mathbf{y}$ *satisfy the conditions from Definition*3.1, *the following identities hold true*:

$$D_\gamma(\mathbf{y};\mathbf{Q}_1) = R^2_{\mathbf{y}\cdot\mathbf{X}_\gamma}, \tag{41}$$

$$D_\gamma(\mathbf{y};\mathbf{Q}_2) = \sum_{i=1}^{k} a_i(\mathbf{y},\mathbf{X}_\gamma) b_i(\mathbf{X}_\gamma), \tag{42}$$

$$D_\gamma(\mathbf{y};\mathbf{Q}_3) = \frac{\sum_{i=1}^{k} r_{i\mathbf{y}}^2}{\prod_{i=2}^{k}[1-R^2_{i\cdot 1,\ldots,(i-1)}]^{1/k}}, \tag{43}$$

*where*

$$a_i(\mathbf{y},\mathbf{X}_\gamma) = R^2_{\mathbf{y}\cdot\mathbf{X}_\gamma} - R^2_{\mathbf{y}\cdot\mathbf{X}_{\gamma\setminus\{i\}}},$$

$$b_i(\mathbf{X}_\gamma) = \frac{\prod_{j=2}^{k-1}[1-R^2_{\varsigma(j)\cdot\varsigma(1),\ldots,\varsigma(j-1)}]^{1/k}}{[1-R^2_{\varsigma(k)\cdot\varsigma(1),\ldots,\varsigma(k-1)}]^{(k-1)/k}},$$

$$\varsigma(j) = \begin{cases} j, & 1 \le j < i, \\ j+1, & i \le j < k, \\ i, & j = k, \end{cases} \tag{44}$$

*and* $r_{i\mathbf{y}}$ *is the correlation between the i-th column of* $\mathbf{X}_\gamma$ *and* $\mathbf{y}$.

See Appendix B for the proof.

Eq. (41) confirms that the interdependence between the columns of $\mathbf{X}_\gamma$ does not have any impact on $D_\gamma(\mathbf{y};\mathbf{Q}_1)$. This is not the case with $D_\gamma(\mathbf{y};\mathbf{Q}_2)$, where the factors $b_i(\mathbf{X}_\gamma)$ measure the linear dependence between the columns of $\mathbf{X}_\gamma$, and they are not affected by the relationship between $\mathbf{y}$ and $\mathbf{X}_\gamma$. Whenever $\mathbf{x}_i$ is a linear combination of some of other columns from $\mathbf{X}_\gamma$, the denominator of $b_i(\mathbf{X}_\gamma)$ becomes zero, whereas the numerator is strictly positive. In this situation, the contribution of $\mathbf{x}_i$ to explaining the variance of $\mathbf{y}$ is marginal, which makes $a_i(\mathbf{y},\mathbf{X}_\gamma)$ to be also zero. From (43), it is evident how multicollinearity affects $D_\gamma(\mathbf{y};\mathbf{Q}_3)$: the denominator goes to zero and the nominator remains strictly positive.

Propositions 3.1 and 3.2 reveal the relationship between the three criteria when the angle $\alpha$ takes extreme values: $\alpha = \pi/2$ and $\alpha \to 0$. It remains open the question on how $\mathrm{SC}_{\rho_2}$ and $\mathrm{SC}_{\rho_3}$ relate to $\mathrm{SC}_{\rho_1}$ when $\alpha \in (0,\pi/2)$. In order to answer the question, we need to make supplementary assumptions on the vector of observations $\mathbf{y}$. This is why we consider next the case of two nested models.

### 3.3. Comparison of the penalty terms when two nested models are tested

Suppose that the model selection problem reduces to deciding if the measurements $\mathbf{y}$ are outcomes from $\mathcal{N}(\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1},\tau\mathbf{I})$ or from $\mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k,\tau\mathbf{I})$, where $\mathbf{X}_{k\text{-}1}$ and $\mathbf{X}_k$ are the same as in Proposition 3.2. The entries of $\boldsymbol{\beta}_{k\text{-}1} \in \mathbb{R}^{k-1}$ and $\boldsymbol{\beta}_k \in \mathbb{R}^k$ are assumed to be non-zero, and $\tau > 0$. After estimating $\hat{\boldsymbol{\beta}}_{k-1}$, $\hat{\boldsymbol{\beta}}_k$ and the noise variance from the available data, one can apply an NML-based criterion to select between the structure $\gamma_{k-1}$ for which the regression matrix is $\mathbf{X}_{k\text{-}1}$, and the structure $\gamma_k$ for which the regression matrix is $\mathbf{X}_k$.

We know from Proposition 3.1 that, disregarding the machinery which has produced $\mathbf{y}$, we have $\mathrm{SC}_{\rho_1}(\mathbf{y};\gamma_{k-1}) = \mathrm{SC}_{\rho_2}(\mathbf{y};\gamma_{k-1}) = \mathrm{SC}_{\rho_3}(\mathbf{y};\gamma_{k-1})$ if $\mathbf{X}_{k\text{-}1}^\top \mathbf{X}_{k\text{-}1} = \mathbf{I}$. Therefore, under the hypothesis of orthonormality for the columns of $\mathbf{X}_{k\text{-}1}$, $D_{\gamma_k}(\mathbf{y};\mathbf{Q}_i)$, $i \in \{1,2,3\}$, is the only term which can potentially make $\mathrm{SC}_{\rho_1}$, $\mathrm{SC}_{\rho_2}$, $\mathrm{SC}_{\rho_3}$ not to take the same decision when choosing between $\gamma_{k-1}$ and $\gamma_k$. To gain more insight, we compute the expectation of $D_{\gamma_k}(\mathbf{y};\mathbf{Q}_i)$ for $i \in \{1,2,3\}$.

**Lemma 3.1.** *If* $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1}, \tau\mathbf{I})$, $\mathbf{X}_{k\text{-}1}^\top \mathbf{X}_{k\text{-}1} = \mathbf{I}$, $\|\mathbf{x}_k\| = 1$ *and* $k > 1$, *then*

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)] = \|\boldsymbol{\beta}_{k\text{-}1}\|^2 + \tau k, \tag{45}$$

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)] = [\|\boldsymbol{\beta}_{k\text{-}1}\|^2 + \tau(k-2+2\omega^{-1})]\omega^{1/k}, \tag{46}$$

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)] = [\|\boldsymbol{\beta}_{k\text{-}1}\|^2 + \tau k + (\mathbf{x}_k^\top \mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1})^2]\omega^{-1/k}, \tag{47}$$

*where* $\mathbb{E}[\cdot]$ *is the expectation operator and* $\omega = \sin^2 \alpha$.

The proof of Lemma 3.1 can be found in Appendix C, where we outline also the proof of the proposition below.

**Proposition 3.4.** *Let* $\varphi_0(\alpha) = 2(1-\sin^{-2}\alpha)/(1-\sin^{-2/k}\alpha)-k$, $\varphi_1(\alpha) = (2-\sin^2\alpha)\sin^{-2/k}\alpha - 1$ *and* $\varphi_2(\alpha) = \sin^{-2/k}\alpha - 1$. *Under the hypotheses of Lemma* 3.1, *we have*:

(a) $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$ *is strictly positive if and only if* $\alpha < \alpha^*$, *where* $\alpha^*$ *is the solution of the equation* $\varphi_0(\alpha) = \|\boldsymbol{\beta}_{k\text{-}1}\|^2/\tau$.

(b) $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$ *takes only non-negative values. Additionally*,

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)] \le \|\boldsymbol{\beta}_{k\text{-}1}\|^2 \varphi_1(\alpha) + \tau k \varphi_2(\alpha). \tag{48}$$

In Proposition 3.4, the Rissanen formula (27) is considered to be a reference, and the other two criteria are compared with it. We see that $\mathrm{SC}_{\rho_2}$ is likely to penalize more than $\mathrm{SC}_{\rho_1}$ the model with structure $\gamma_k$ only when the angle between $\langle \mathbf{X}_{k\text{-}1} \rangle$ and $\langle \mathbf{x}_k \rangle$ is smaller than a threshold. The value of the threshold is mainly given by the ratio $\|\boldsymbol{\beta}_{k\text{-}1}\|^2/\tau$, which in our case equals the energy-to-noise ratio (ENR) because $\|\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1}\| = \|\boldsymbol{\beta}_{k\text{-}1}\|$. Since $\lim_{\alpha\to 0}\varphi_0(\alpha) = \infty$ and $\lim_{\alpha\to\pi/2}\varphi_0(\alpha) = k$, the solution $\alpha^*$ is guaranteed to exist when ENR $> k$. Moreover, if ENR is larger than $k$, then $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$ attains its minimum when the principal angle takes value $\alpha_{\min} = \arcsin(\omega_{\min}^{1/2})$, where $\omega_{\min} = 2(k-1)/(\|\boldsymbol{\beta}_{k\text{-}1}\|^2/\tau + (k-2))$. The increase of ENR makes $\alpha_{\min}$ to decrease and $\alpha^*$ to be closer to zero such that $\mathrm{SC}_{\rho_2}$ penalizes more severely than $\mathrm{SC}_{\rho_1}$ only when $\alpha \approx 0$.

On contrary, $\mathrm{SC}_{\rho_3}$ penalizes the $\gamma_k$-model more stringently than $\mathrm{SC}_{\rho_1}$ for all $\alpha \in (0,\pi/2)$. Observe in (48) that $\varphi_1(\alpha)$ and $\varphi_2(\alpha)$ are monotonically decreasing functions, and the upper bound for $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$ goes down from $\infty$ to zero when $\alpha$ increases from zero to $\pi/2$.

To complete the analysis, we provide the analogue of Proposition 3.4 for the case when $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k, \tau\mathbf{I})$. Let us assume that the eigenvalues of $\mathbf{X}_k^\top \mathbf{X}_k$ are $\lambda_1, \ldots, \lambda_k$, and all of them are strictly positive. To write more compactly the results, we define: $\mathcal{A}_\lambda = (\sum_{i=1}^k \lambda_i)/k$ (arithmetic mean), $\mathcal{G}_\lambda = (\prod_{i=1}^k \lambda_i)^{1/k}$ (geometric mean) and $\mathcal{H}_\lambda = k/\sum_{i=1}^k \lambda_i^{-1}$ (harmonic mean). The expressions of $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_i)]$ for $i \in \{1,2,3\}$ are given in the lemma below.

**Lemma 3.2.** *If* $\mathbf{y} \sim \mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k, \tau\mathbf{I})$, $\mathbf{X}_{k\text{-}1}^\top \mathbf{X}_{k\text{-}1} = \mathbf{I}$, $\|\mathbf{x}_k\| = 1$ *and* $k > 1$, *then*

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)] = \|\mathbf{X}_k\boldsymbol{\beta}_k\|^2 + \tau k, \tag{49}$$

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)] = \|\boldsymbol{\beta}_k\|^2 \mathcal{G}_\lambda + \tau k(\mathcal{G}_\lambda/\mathcal{H}_\lambda), \tag{50}$$

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)] = [\boldsymbol{\beta}_k^\top (\mathbf{X}_k^\top \mathbf{X}_k)^2 \boldsymbol{\beta}_k]/\mathcal{G}_\lambda + \tau k(\mathcal{A}_\lambda/\mathcal{G}_\lambda). \tag{51}$$

**Proof.** The results are easily obtained by applying the formula of the expectation for quadratic forms [30, p. 439]. □

Lemma 3.2 helps us to find bounds for $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$ and $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$, which are similar with those given in Proposition 3.4.

**Proposition 3.5.** *Under the hypotheses of Lemma* 3.2, *we have*

(a) *Let* $\psi_1(\alpha) = \sin^{2/k}\alpha - \cos\alpha - 1$, $\psi_2(\alpha) = \sin^{2/k}\alpha + \cos\alpha - 1$ *and* $\psi_3(\alpha) = \sin^{2/k}\alpha/(1-\cos\alpha) - 1$. *Then*

$$\|\boldsymbol{\beta}_k\|^2 \psi_1(\alpha) \le \mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)] \le \|\boldsymbol{\beta}_k\|^2 \psi_2(\alpha) + \tau k \psi_3(\alpha). \tag{52}$$

(b) *Let* $\psi_4(\alpha) = -\sin^{2/k}\alpha/4$,
$\psi_5(\alpha) = (1-\cos\alpha)^2/\sin^{2/k}\alpha + \cos\alpha - 1$,
$\psi_6(\alpha) = (1+\cos\alpha)^2/\sin^{2/k}\alpha - \cos\alpha - 1$
*and* $\psi_7(\alpha) = (1+\cos\alpha)/\sin^{2/k}\alpha - 1$. *For* $\alpha \in (0,\pi/2]$,

$$\|\boldsymbol{\beta}_k\|^2 \psi_4(\alpha) \le \mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)], \tag{53}$$

*and for* $\alpha \in [\pi/3,\pi/2]$, *the inequality becomes*

$$\|\boldsymbol{\beta}_k\|^2 \psi_4(\alpha) \le \|\boldsymbol{\beta}_k\|^2 \psi_5(\alpha) \le \mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]. \tag{54}$$

*Additionally*,

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)] \le \|\boldsymbol{\beta}_k\|^2 \psi_6(\alpha) + \tau k \psi_7(\alpha), \tag{55}$$

*for all* $\alpha \in (0,\pi/2]$.

Proof is deferred to Appendix C.

Note in (52) that the span of $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2) - D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$ is given by $\|\boldsymbol{\beta}_k\|^2[\psi_2(\alpha) - \psi_1(\alpha)]$ and $\tau k \psi_3(\alpha)$. The second term is the dominant one when $\alpha$ is close to zero, as we can see from $\lim_{\alpha\to 0}\{\|\boldsymbol{\beta}_k\|^2[\psi_2(\alpha) - \psi_1(\alpha)]\} = 2\|\boldsymbol{\beta}_k\|^2 < \infty$ and $\lim_{\alpha\to 0}\{\tau k \psi_3(\alpha)\} = \infty$. To monitor the decrease of the two terms when $\alpha$ varies from zero to $\pi/2$, we define $\mathfrak{F}_{\psi_1,\psi_2}(\alpha_1,\alpha_2) = (\psi_2(\alpha_2) - \psi_1(\alpha_2))/(\psi_2(\alpha_1) - \psi_1(\alpha_1))$ and $\mathfrak{F}_{\psi_3}(\alpha_1,\ \alpha_2) = \ \psi_3(\alpha_2)/\psi_3(\alpha_1)$, where $0 < \alpha_1 < \alpha_2 < \pi/2$. For example, when $k$=6, we get $\mathfrak{F}_{\psi_1,\psi_2}(\pi/180,\pi/6) \approx 87\%$, $\mathfrak{F}_{\psi_1,\psi_2}\,(\pi/6,\pi/3) \approx 58\%$ and $\mathfrak{F}_{\psi_1,\psi_2}(\pi/3,\pi/2-\pi/180) \approx 3\%$, whereas $\mathfrak{F}_{\psi_3}(\pi/180,\pi/6) \approx 0.3\%$, $\mathfrak{F}_{\psi_3}(\pi/6,\pi/3) \approx 18\%$ and $\mathfrak{F}_{\psi_3}(\pi/3,\pi/2-\pi/180) \approx 2\%$. Remark that the term given by $\psi_3(\cdot)$ diminishes significantly when $\alpha$ increases from $\pi/180$ to $\pi/6$. Another significant reduction occurs for both terms in the interval $[\pi/6,\pi/2-\pi/180]$. An important observation is that the upper bound in (52) increases monotonically with $\tau$ when $\mathbf{X}_k$ and $\boldsymbol{\beta}_k$ are fixed. Therefore, when the ENR lowers, there exists a higher chance that $\mathrm{SC}_{\rho_2}$ penalizes the $\gamma_k$-model more stringently than $\mathrm{SC}_{\rho_1}$. This finding is of special interest because, in Proposition 3.5, the model with structure $\gamma_k$ is assumed to be the "true" one.

A similar analysis can be done for $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$. In the vicinity of zero, the span of $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$ is given by $\|\boldsymbol{\beta}_k\|^2[\psi_6(\alpha)-\psi_4(\alpha)]$ and $\tau k\psi_7(\alpha)$. Since $\lim_{\alpha\to 0}\{\|\boldsymbol{\beta}_k\|^2[\psi_6(\alpha)-\psi_4(\alpha)]\} = \lim_{\alpha\to 0}\{\tau k\psi_7(\alpha)\}=\infty$, the two terms are equally important, not like in the case of $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$. It is also interesting to point out for $k=6$ that $(\psi_6(\pi/2-\pi/180)-\psi_5(\pi/2-\pi/180))/(\psi_6(\pi/3)-\psi_5(\pi/3)) \approx \psi_7(\pi/2-\pi/180)/\psi_7(\pi/3)\approx 3\%$, which is similar with the result found previously for $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$.

## 4. Experimental results

### 4.1. Model selection criteria used in experiments

We illustrate the performance of $\mathrm{SC}_{\rho_1}$, $\mathrm{SC}_{\rho_2}$ and $\mathrm{SC}_{\rho_3}$ against other criteria. For ease of comparison, we employ for all the model selection rules the same notations like those from (3)–(4). As already told in Section 1, BIC is among the most popular criteria, and this is why we include it in our experiments. The well-known expression of BIC is [29]

$$\mathrm{BIC}(\mathbf{y};\gamma)=\frac{n}{2}\ln\hat{\tau}_\gamma+\frac{k}{2}\ln n. \tag{56}$$

Another widely used criteria are AIC and its bias corrected version which is called $\mathrm{AIC_c}$ [13]. Recently, Seghouane has applied bootstrap-type techniques to obtain $\mathrm{AIC_{c3}}$, a new corrected version of AIC. The complete derivation can be found in [32], where it was also shown experimentally that, for small sample size, $\mathrm{AIC_{c3}}$ outperforms $\mathrm{AIC_c}$ as well as two other corrected criteria: $\mathrm{AIC_c^*}$ [33] and $\mathrm{KIC_c}$ [34]. Remark that the small sample size case makes the difference between various forms of AIC because asymptotically all of them are equivalent. For the sake of comparison, we consider in our simulations the criterion from [32]:

$$\mathrm{AIC_{c3}}(\mathbf{y};\gamma)=\frac{n}{2}\ln\hat{\tau}_\gamma+\frac{(k+1)(n+k+2)}{n-k-2}-\frac{k}{n-k}. \tag{57}$$

Following the suggestion of one of the reviewers, we briefly discuss how $\mathrm{SC}_{\rho_1}$ relates to BIC and AIC. The aim of the discussion is to provide support for the interpretation of the experimental results presented in this section. Note that the formula of $\mathrm{SC}_{\rho_1}$ from (27) can be re-written as follows [7, Eq. (16)]:

$$\frac{1}{2}\mathrm{SC}_{\rho_1}(\mathbf{y};\gamma)=\frac{n}{2}\ln\hat{\tau}_\gamma+\frac{k}{2}\ln F_\gamma+\frac{1}{2}\ln\frac{k}{(n-k)^{n-1}}, \tag{58}$$

where $F_\gamma=(\|\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2/(nk))/(\hat{\tau}_\gamma/(n-k))$. It is evident that the goodness-of-fit term is the same for all the criteria in (56)–(58). The key difference is that $F_\gamma$ from (58) depends on the data vector $\mathbf{y}$, while the penalty terms from (56) and (57) depend only on $n$ and $k$. Let us observe that $F_\gamma$ coincides with the $F$-statistic which is used to test the hypothesis that each entry of $\hat{\boldsymbol{\beta}}_\gamma$ is zero [17, Section 5; 31, Chapter 4].

More importantly, by applying the settings from [4], it was worked out in [10] an expression of $F_\gamma$ which leads to the conclusion that, asymptotically, $\mathrm{SC}_{\rho_1}$ combines the strengths of both BIC and AIC. Similarly with BIC, $\mathrm{SC}_{\rho_1}$ is *consistent*: if the "true model" is finite-dimensional and is included in the set of candidates, then the probability that this model is selected goes to one as the sample size increases [8]. However, if the "true model" is not finite-dimensional, then $\mathrm{SC}_{\rho_1}$ is asymptotically *efficient* in the sense that selects the candidate model which minimizes the one-step mean squared error of prediction. The same property has been proved for AIC long time ago [35]. We refer to [10] for the technical details concerning the results outlined above.

The two-part MDL criterion, which is equivalent to BIC, was refined in [22] such that its penalty term involves the logarithm of determinant of the observed FIM. A similar formula, which is not rooted in information theory, was proposed by Kay [15]:

$$\mathrm{CME}(\mathbf{y};\gamma)=\frac{n-k-2}{2}\ln\frac{n\hat{\tau}_\gamma}{n-k}+\frac{1}{2}\ln|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|+\ln\frac{[\pi(n-k)]^{(n-k)/2}}{\Gamma\left(\frac{n-k}{2}\right)}. \tag{59}$$

The significance of the acronym CME is conditional model estimator.

In addition to BIC, $\mathrm{AIC_{c3}}$ and CME, we include in our tests the $\mathrm{MML_g}$ criterion from [28]:

$$\mathrm{MML_g}(\mathbf{y};\gamma)=\frac{n-k+2}{2}\left(\ln\frac{n\hat{\tau}_\gamma}{n-k+2}+1\right)+\frac{k-2}{2}\ln\frac{\|\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2}{\max\{k-2,1\}}+\frac{1}{2}\ln[(n-k)k^2].$$

The formula above is applied whenever $\|\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma\|^2/\max\{k-2,1\}>n\hat{\tau}_\gamma/(n-k+2)$ and $k>0$. Otherwise, it is used as follows:

$$\mathrm{MML_g}(\mathbf{y};\emptyset)=\frac{n}{2}\left(\ln\frac{n\hat{\tau}_\gamma}{n-k+2}+1\right)+\frac{1}{2}\ln(n-1)+\frac{1}{2}.$$

For completeness, we also consider a second criterion from [28]:

$$\mathrm{MML_u}(\mathbf{y};\gamma)=\frac{n-k}{2}\ln(2\pi)+\frac{n-k}{2}\left(\ln\frac{n\hat{\tau}_\gamma}{n-k}+1\right)+\frac{k}{2}\ln(\pi\mathbf{y}^\top\mathbf{y})-\ln\Gamma\left(\frac{k}{2}+1\right)+\frac{1}{2}\ln(k+1).$$

Remark that the expression above is for both $k=0$ and $k>0$.

Next we conduct experiments for simulated and real life data sets.

### 4.2. Numerical examples

*Example* 1 illustrates the case of two nested models, which is akin to the model selection problem discussed in Section 3.3. We generate randomly $k$ vectors $\mathbf{z}_1,\ldots,\mathbf{z}_k\in\mathbb{R}^{n\times 1}$ such that $\mathbf{z}_i^\top\mathbf{z}_j=\delta_{i,j}$ for all $i,j\in\{1,\ldots,k\}$, with the convention that $\delta_{\cdot,\cdot}$ denotes the Kronecker operator. In our settings, $k=6$ and $n=50$. Then we choose $\alpha\in(0,\pi/2]$, and define the matrices $\mathbf{X}_{k\text{-}1}=[\mathbf{z}_1\cdots\mathbf{z}_{k-1}]$ and $\mathbf{X}_k=[\mathbf{X}_{k\text{-}1}\ \mathbf{x}_k]$, where $\mathbf{x}_k = \mathbf{z}_1\cos(\alpha)+\mathbf{z}_k\sin(\alpha)$. It is evident that

$$\mathbf{X}_{k\text{-}1}^\top \mathbf{X}_{k\text{-}1} = \mathbf{I}, \tag{60}$$

$$\mathbf{X}_k^\top \mathbf{X}_k = \begin{bmatrix} 1 & & \cos\alpha \\ & \ddots & \\ \cos\alpha & & 1 \end{bmatrix}. \tag{61}$$

More importantly, $\alpha$ is the principal angle between the subspaces $\langle \mathbf{X}_{k-1} \rangle$ and $\langle \mathbf{x}_k \rangle$. Given $\alpha$, we aim to test the performance of various criteria in deciding if the observations are outcomes from $\mathcal{N}(\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1},\tau\mathbf{I})$ or from $\mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k,\tau\mathbf{I})$. Therefore, we simulate the measurements as follows:

- In the first scenario, we take $\mathbf{y} = \mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1}+\sqrt{\tau}\mathbf{d}$, where $\boldsymbol{\beta}_{k\text{-}1} = [1\ldots 1]^\top$, $\tau = (k-1)/n$ and $\mathbf{d} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$.
- In the second scenario, we have $\mathbf{y} = \mathbf{X}_k\boldsymbol{\beta}_k+\sqrt{\tau}\mathbf{d}$, where $\boldsymbol{\beta}_k = [1\ldots 1]^\top$, $\tau = (k+2\cos\alpha)/n$ and $\mathbf{d}$ has the same significance as above.

Based on (60), the signal-to-noise ratio in the first case is given by $\text{SNR} = \|\mathbf{X}_{k-1}\boldsymbol{\beta}_{k-1}\|^2/(n\tau) = \|\boldsymbol{\beta}_{k-1}\|^2/(n\tau) = 1$. Similarly, by using (61) for the second case, we get

$$\text{SNR} = \frac{\|\mathbf{X}_k\boldsymbol{\beta}_k\|^2}{n\tau} = \frac{k+2\cos\alpha}{n\tau} = 1. \tag{62}$$

For each $\alpha \in \{\pi/180, 2\pi/180, \ldots, \pi/2\}$, we generate randomly 500 different realizations of the matrix $\mathbf{X}_k$ by applying the procedure described above. The first $k-1$ columns of each $\mathbf{X}_k$-matrix define the corresponding $\mathbf{X}_{k\text{-}1}$-matrix. Furthermore, every $\mathbf{X}_{k\text{-}1}$-matrix is used to yield 500 $\mathbf{y}$-vectors, according to the first scenario. Hence, for each angle $\alpha$, we have $25\times 10^4$ data vectors which are outcomes from $\mathcal{N}(\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1},\tau\mathbf{I})$. Then we decide for each $\mathbf{y}$ if the best model structure is $\gamma_{k-1}$ or $\gamma_k$ by employing the eight criteria whose performance is evaluated. In Fig. 1 is plotted the empirical probability of correct estimation versus the angle $\alpha$. A similar experiment is done for $25\times 10^4$ data vectors simulated, for each $\alpha$, according to the second scenario. The estimation results are shown in Fig. 2.

In Figs. 1 and 2, we also plot the normalized condition number for the matrix $\mathbf{X}_k$: $\text{ncond}(\alpha) = \text{cond}(\alpha)/\text{cond}(\alpha_0)$, where $\alpha_0 = \pi/180$. For an arbitrary $\alpha$, $\text{cond}(\alpha)$ denotes the 2-norm condition number of $\mathbf{X}_k$, and it equals $[\lambda_{\max}(\alpha)/\lambda_{\min}(\alpha)]^{1/2}$, where $\lambda_{\max}(\alpha)$ and $\lambda_{\min}(\alpha)$ are the maximum and the minimum eigenvalues of the matrix $\mathbf{X}_k^\top\mathbf{X}_k$ [30, p. 78]. It is clear that, for $\alpha$ close to zero, $\mathbf{X}_k$ is badly conditioned numerically. For instance, $\text{cond}(\alpha_0) \approx 115$. However, $\text{cond}(\alpha)$ becomes rapidly smaller when $\alpha$ increases, and we mark in Figs. 1 and 2 the point that corresponds to the value 10 of the 2-norm condition number.

Observe in Fig. 1 that, for all $\alpha$, $\text{SC}_{\rho_1}$ selects the true model with high probability. The fact that the performance of $\text{SC}_{\rho_1}$ is not affected by the geometry of the linear subspaces $\langle \mathbf{X}_{k\text{-}1} \rangle$ and $\langle \mathbf{x}_k \rangle$ is in line with the results from Section 3.3 (see, for example, Eq. (45) in Lemma 3.1). We remark also in Fig. 1 that the behavior of $\text{MML}_\text{g}$, $\text{MML}_\text{u}$, BIC and $\text{AIC}_\text{c3}$ is very similar with that of $\text{SC}_{\rho_1}$.

The relationship between the performance of $\text{SC}_{\rho_2}$ and $\text{SC}_{\rho_1}$ can be understood better by recalling that, according to Proposition 3.4, the difference of the expectations of penalty terms, $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)]-\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$, is positive only for $\alpha \in (0,\alpha^*)$, it decreases as long as $\alpha \le \alpha_{\min}$, and then increases when $\alpha \in (\alpha_{\min},\pi/2)$. This is very well reflected by the graphs within Fig. 1, where $\text{SC}_{\rho_2}$ is slightly better than $\text{SC}_{\rho_1}$ when $\alpha = \pi/180$, but its performance declines when $\alpha$ increases and, after reaching a minimum point, $\text{SC}_{\rho_2}$ improves such that it becomes as good as $\text{SC}_{\rho_1}$ when $\alpha = \pi/2$.

From the identities (49) and (62), we get $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)] = (k+2\cos\alpha)(1+k/n)$. So, we expect that, with our experimental settings, the criterion $\text{SC}_{\rho_1}$ penalizes less the $\gamma_k$-model when $\alpha$ increases. This theoretical result, which is based on Lemma 3.2, agrees perfectly with the empirical results shown in Fig. 2.

By looking simultaneously at Figs. 1 and 2, we note that $\text{SC}_{\rho_2}$ prefers the $\gamma_k$-model when the condition number takes large values, and this effect is undesirable. On contrary, $\text{SC}_{\rho_3}$ selects the $\gamma_{k-1}$-model whenever the condition number is high, which shows that $\text{SC}_{\rho_3}$ is prone to choose the model whose explanatory variables are linearly independent, and not the "true" model. When $\alpha = \pi/2$, or equivalently the matrix $\mathbf{X}_k$ is orthonormal, the criteria $\text{SC}_{\rho_1}$, $\text{SC}_{\rho_2}$ and $\text{SC}_{\rho_3}$ reduce to one single criterion, as we know from Proposition 3.1.

In Fig. 1, CME has the poorest results as it strongly prefers the $\gamma_k$-model. This can be explained by noticing in (59) that $\frac{1}{2}\ln|\mathbf{X}_{k\text{-}1}^\top\mathbf{X}_{k\text{-}1}|$ is a penalty term for the $\gamma_{k-1}$-model, and $\frac{1}{2}\ln|\mathbf{X}_k^\top\mathbf{X}_k|$ is a penalty term for the $\gamma_k$-model. In our settings, $\frac{1}{2}\ln|\mathbf{X}_{k\text{-}1}^\top\mathbf{X}_{k\text{-}1}| = 0$, whereas $\frac{1}{2}\ln|\mathbf{X}_k^\top\mathbf{X}_k| \to -\infty$ when $\alpha \to 0$. It is worth mentioning that $\ln|\mathbf{X}_k^\top\mathbf{X}_k|$ is also a penalty term within $\text{SC}_{\rho_2}$-formula in (31). However, the significant decrease of $\ln|\mathbf{X}_k^\top\mathbf{X}_k|$ when $\alpha \to 0$ is compensated in $\text{SC}_{\rho_2}$-formula by the increase of the term $k\ln(\|\hat{\boldsymbol{\beta}}_k\|^2/n/k)$. More interestingly, CME has difficulties in correctly identifying the $\gamma_{k-1}$-model even when $\alpha$ takes values close to $\pi/2$. The reason is that the logarithm of determinant of the observed FIM is not guaranteed to be a correct penalty term even if the columns of $\mathbf{X}$ are almost orthogonal. We will investigate more carefully this aspect in the next example.

*Example* 2 is taken from [16] and is focused on the variable selection for the linear regression in (1), when the matrix $\mathbf{X}$ has the particular form

$$\mathbf{X} = \begin{bmatrix} \cos(2\pi f_1) & \cdots & \cos(2\pi f_8) \\ \vdots & \ddots & \vdots \\ \cos[2\pi f_1(N-1)] & \cdots & \cos[2\pi f_8(N-1)] \end{bmatrix},$$

where $f_j = [0.10+(j-1)/100]$ for $j \in \{1,\ldots,8\}$. With the notations from (1), $n = N-1$ and $m = 8$. The vector of linear parameters $\boldsymbol{\beta}$ contains the unknown amplitudes, and the variance of the additive Gaussian noise is assumed to be unknown. The competitors are eight nested models with structures $\gamma_1,\ldots,\gamma_8$, where $\gamma_k = \{1,\ldots,k\}$. Equivalently, the regressor matrix $\mathbf{X}_{\gamma_k}$ for the model $\gamma_k$, $k \in \{1,\ldots,8\}$, is given by the first $k$ columns of $\mathbf{X}$. For simplicity, we use the notation $\mathbf{X}_k$ instead of $\mathbf{X}_{\gamma_k}$, and $\boldsymbol{\beta}_k$ instead of $\boldsymbol{\beta}_{\gamma_k}$.

To mimic the experiments from [16], we simulate data according to the structure $\gamma_3$ by tacking $\boldsymbol{\beta}_3 = [1\ 1\ 1]^\top$. In the first experiment, the noise variance is $\tau = 10$ and the sample size is varied by choosing $N$ from the set $\{100,110,\ldots,300\}$. In the second experiment, the sample size is kept fixed

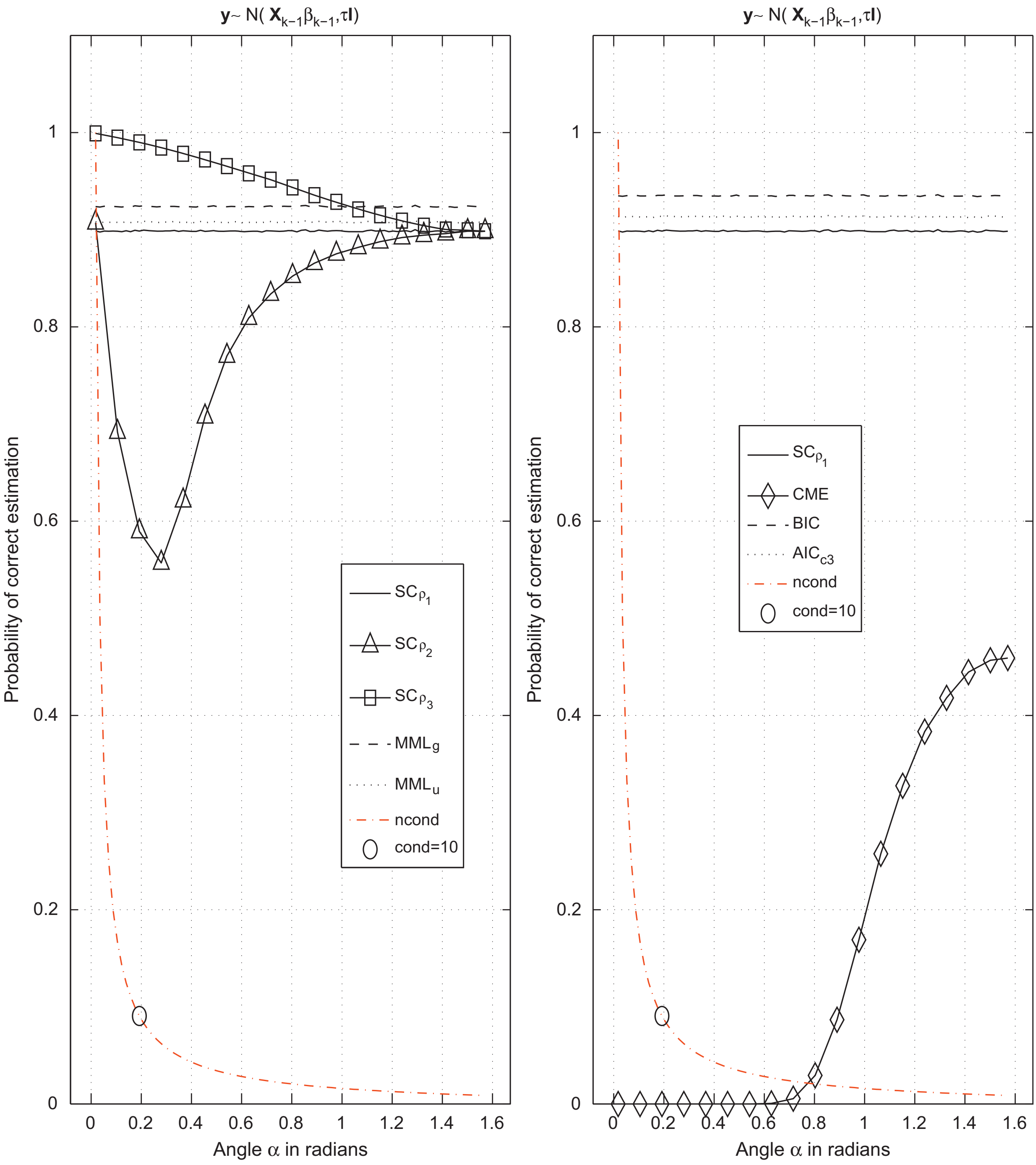

**Fig. 1.** Example 1—the empirical probability of deciding correctly that the observations $\mathbf{y} \in \mathbb{R}^n$ are outcomes from $\mathcal{N}(\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1},\tau\mathbf{I})$, and not from $\mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k,\tau\mathbf{I})$. With the convention that $\mathbf{X}_k = [\mathbf{X}_{k\text{-}1}\ \mathbf{x}_k]$, $\alpha$ denotes the principal angle between the linear subspaces $\langle \mathbf{X}_{k\text{-}1} \rangle$ and $\langle \mathbf{x}_k \rangle$. For an arbitrary $\alpha$, cond($\alpha$) denotes the 2-norm condition number of $\mathbf{X}_k$. The normalized condition number is ncond($\alpha$) = cond($\alpha$)/cond($\alpha_0$), where $\alpha_0 = \pi/180$. For the simulated data, $n = 50$, $k = 6$, and $\tau$ is chosen such that SNR = 0 dB.

($N$=100), and the SNR is varied by modifying the noise variance such that $1/\tau \in \{0.01, 0.02, \ldots, 0.2\}$. The empirical probabilities of selecting correctly the number of sinusoids are plotted in Fig. 3 for the first experiment, and in Fig. 4 for the second experiment. Note that the probabilities shown in Fig. 3 are obtained, for each value of $N$, from $10^4$ runs. Similarly, in the second experiment, the number of runs for each value of $1/\tau$ is $10^4$.

In both figures, the graphs for $\mathrm{SC}_{\rho_1}$, $\mathrm{SC}_{\rho_2}$ and $\mathrm{SC}_{\rho_3}$ almost coincide. This is because [5,16,14]

$$\mathbf{X}_k^\top \mathbf{X}_k \approx (n/2)\mathbf{I}, \quad \forall k \in \{2, \ldots, 8\}, \qquad (63)$$

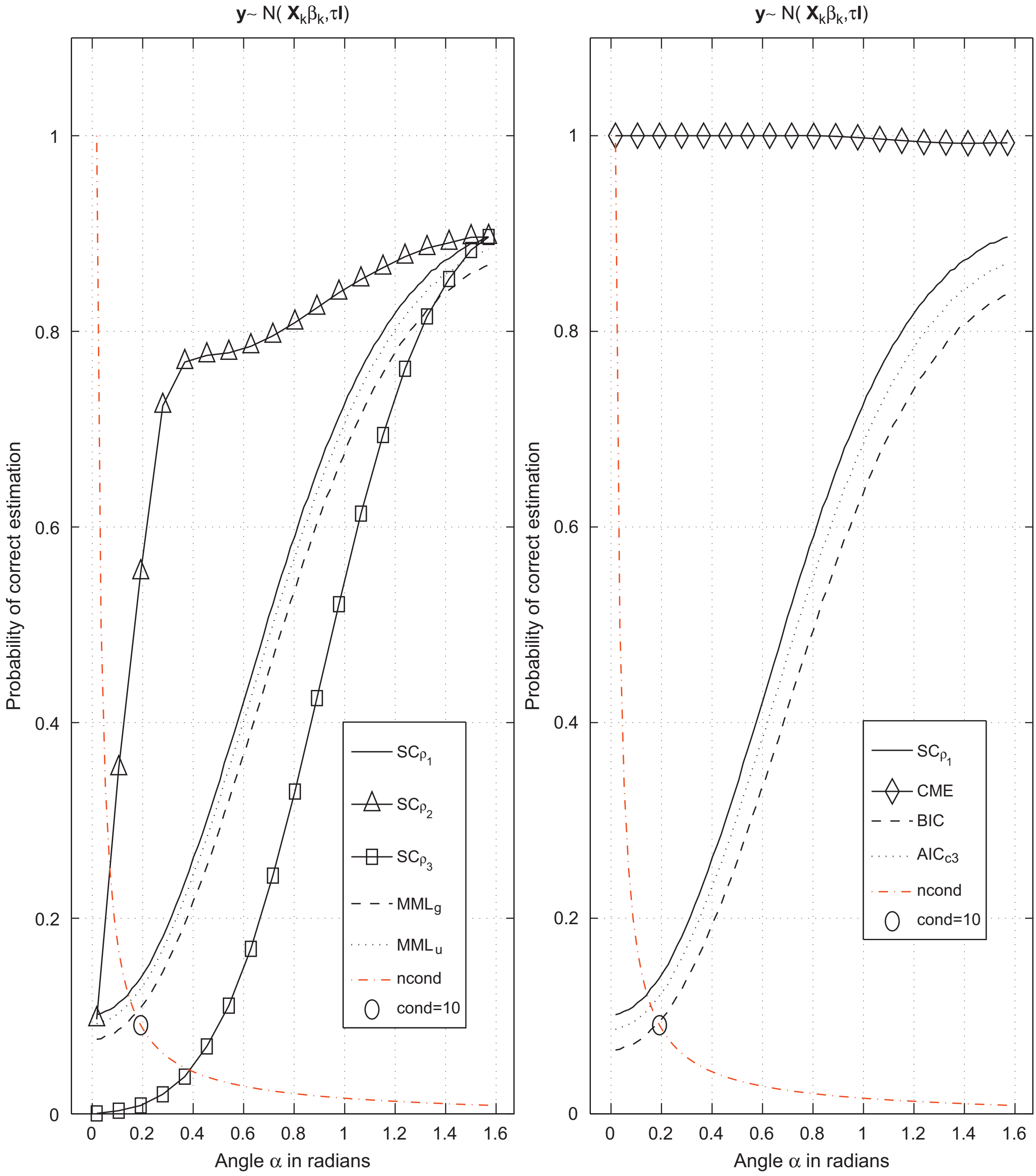

**Fig. 2.** Example 1—the empirical probability of deciding correctly that the observations $\mathbf{y} \in \mathbb{R}^n$ are outcomes from $\mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k,\tau\mathbf{I})$, and not from $\mathcal{N}(\mathbf{X}_{k-1}\boldsymbol{\beta}_{k-1},\tau\mathbf{I})$. All conventions are the same like in Fig. 1.

which makes the condition within point (a) of Proposition 3.1 to be satisfied. Additionally, MML$_g$ and MML$_u$ perform similarly with SC$_{\rho_1}$, and they are both superior to SC$_{\rho_1}$ only when $N>200$ as we can see in Fig. 3. We observe in the same figure that AIC$_{c3}$ outperforms other criteria when $N<150$. The good estimation capabilities of AIC$_{c3}$ when sample size is small can be noticed also in Fig. 4 where, for $N=100$, AIC$_{c3}$ is superior to SC$_{\rho_1}$ and BIC for almost all SNRs. On contrary, when $N>200$, the estimation results of AIC$_{c3}$ are modest, and BIC improves significantly. The reason is simple: AIC$_{c3}$ has been designed especially for the small sample case [32], whereas the use of BIC is recommended for large samples because its derivation relies on asymptotic approximations [29]. It is remarkable that SC$_{\rho_1}$ is nearly as

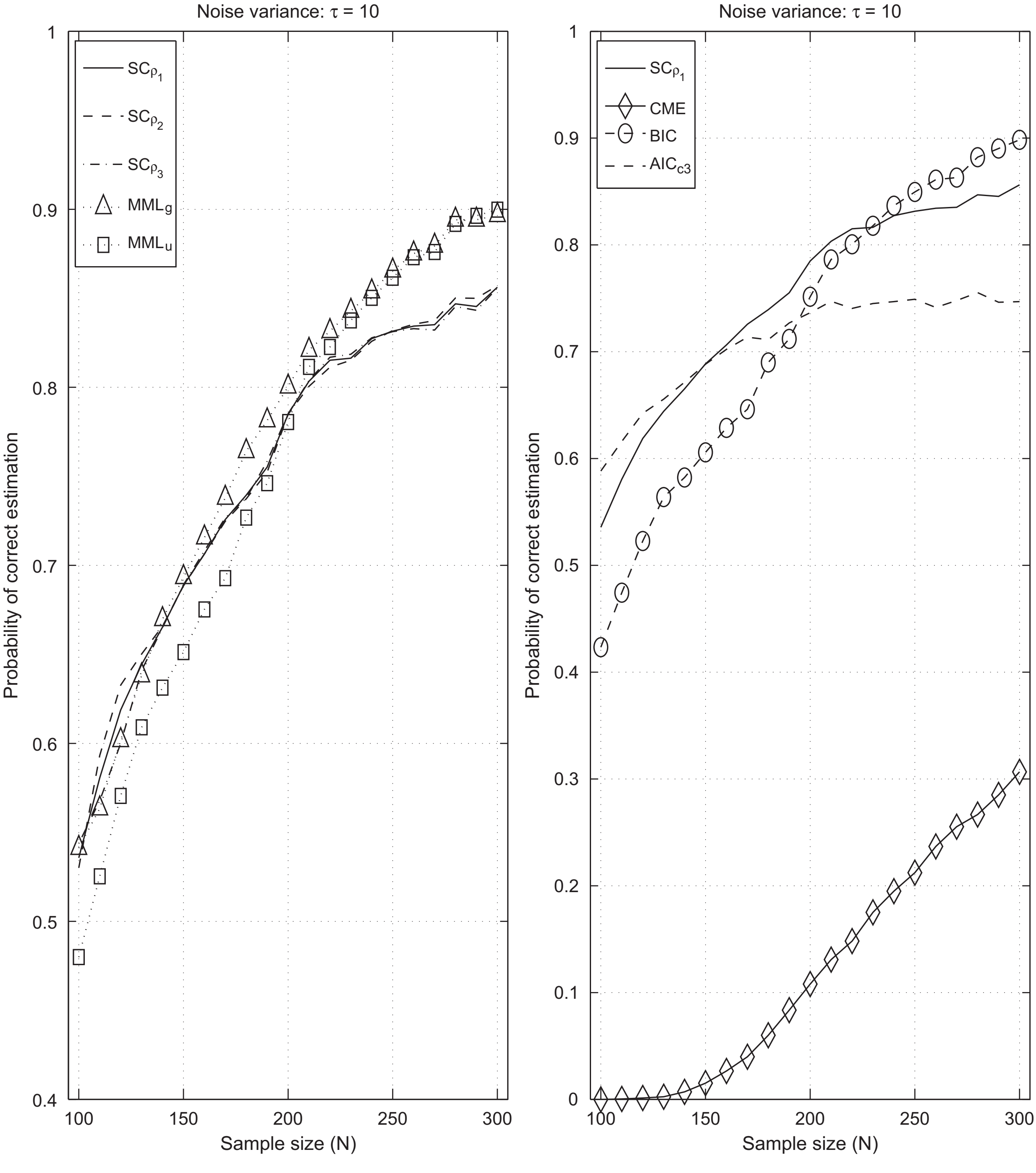

**Fig. 3.** Example 2—the empirical probability of estimating correctly the number of sinusoids versus the sample size. Note that the range of values being presented along the vertical axes is different for the two plots.

good as $AIC_{c3}$ when $N$ is small, and it is only marginally inferior to BIC when $N$ is large.

The performance of CME is again very modest, and it can be explained by re-writing, in a more convenient form, the expression from (59). We approximate $\ln\Gamma((n-k)/2)$ by (26), and then we neglect the sum $(n/2)\ln[2\pi n\exp(1)]-\frac{1}{2}\ln(4\pi n^2)$ which does not depend on $k$. So, we obtain the following formula when the structure is $\gamma_k$:

$$\mathrm{CME}(\mathbf{y};\gamma_k)=\frac{n}{2}\ln\hat{\tau}_k+\frac{1}{2}\ln\left|\frac{(n-k)/n}{2\pi\exp(1)\hat{\tau}_k}\mathbf{X}_k^\top\mathbf{X}_k\right| -\left[\ln\hat{\tau}_k+\frac{n-3}{2}\ln(n-k)\right]. \quad (64)$$

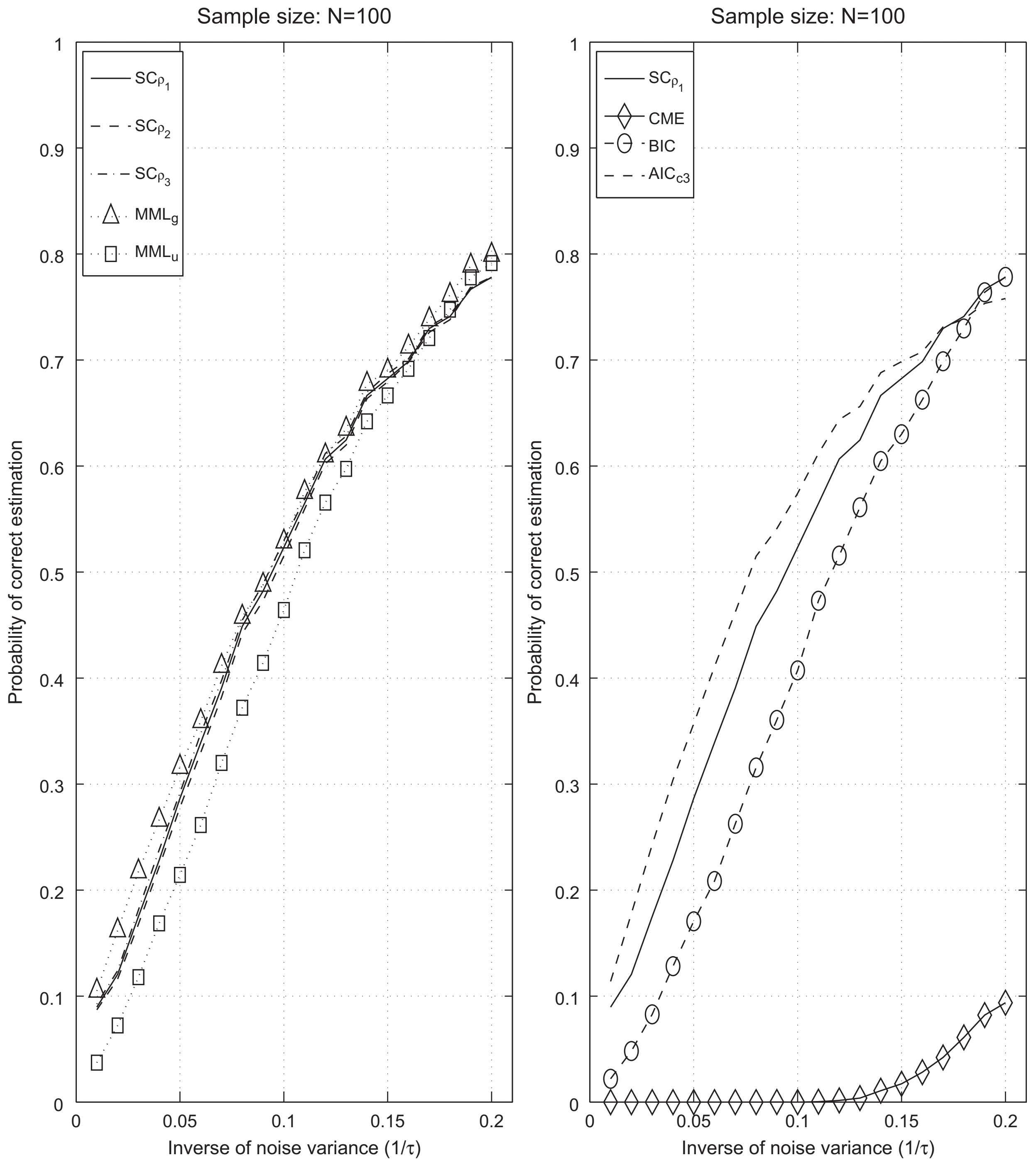

**Fig. 4.** Example 2—the empirical probability of estimating correctly the number of sinusoids versus the inverse of the noise variance.

It is obvious that $\hat{\tau}_k = \|(\mathbf{I}-\mathbf{P}_k)\mathbf{y}\|^2/n$, where $\mathbf{P}_k$ is the orthogonal projection matrix onto the linear subspace $\langle \mathbf{X}_k \rangle$. By using (63), we notice that the second term within (64) is given by $\text{PEN}(\mathbf{y};\gamma_k) = (k/2)\ln[(n-k)/(4\pi\ \exp(1)\hat{\tau}_k)]$. For small $n$, $\text{PEN}(\mathbf{y};\gamma_k)$does not increase fast enough when the model order $k$ becomes larger. For comparison, note in (56) that the BIC penalty term is $(k/2)\ln n$. The fact that the penalty of CME is possibly incorrect for small sample size has been already analyzed in the case when the variance of the Gaussian noise is a priori known (see [16]). However, we show in the next example that CME is rather good in estimating the order of a polynomial model.

*Example* 3 is also taken from [16], and this time the regressor matrix is given by

$$\mathbf{X}=\begin{bmatrix} 1^0 & 1^1 & \cdots & 1^9 \\ 2^0 & 2^1 & \cdots & 2^9 \\ \vdots & \vdots & \ddots & \vdots \\ (N-1)^0 & (N-1)^1 & \cdots & (N-1)^9 \end{bmatrix}.$$

It is evident that $n=N-1$ and $m=10$. Similarly with the previous example, the number of competing nested models equals $m$ and their structures are such that $\gamma_k=\{1,\ldots,k\}$ for all $k\in\{1,\ldots,m\}$. The variance of the additive Gaussian noise is assumed to be unknown, and we use again the notation $\mathbf{X}_k$ instead of $\mathbf{X}_{\gamma_k}$, and $\boldsymbol{\beta}_k$ instead of $\boldsymbol{\beta}_{\gamma_k}$.

The data are simulated according to the structure $\gamma_3$ such that the linear parameters are $\boldsymbol{\beta}_3=[0\ 0.4\ 0.1]^\top$. Hence, the

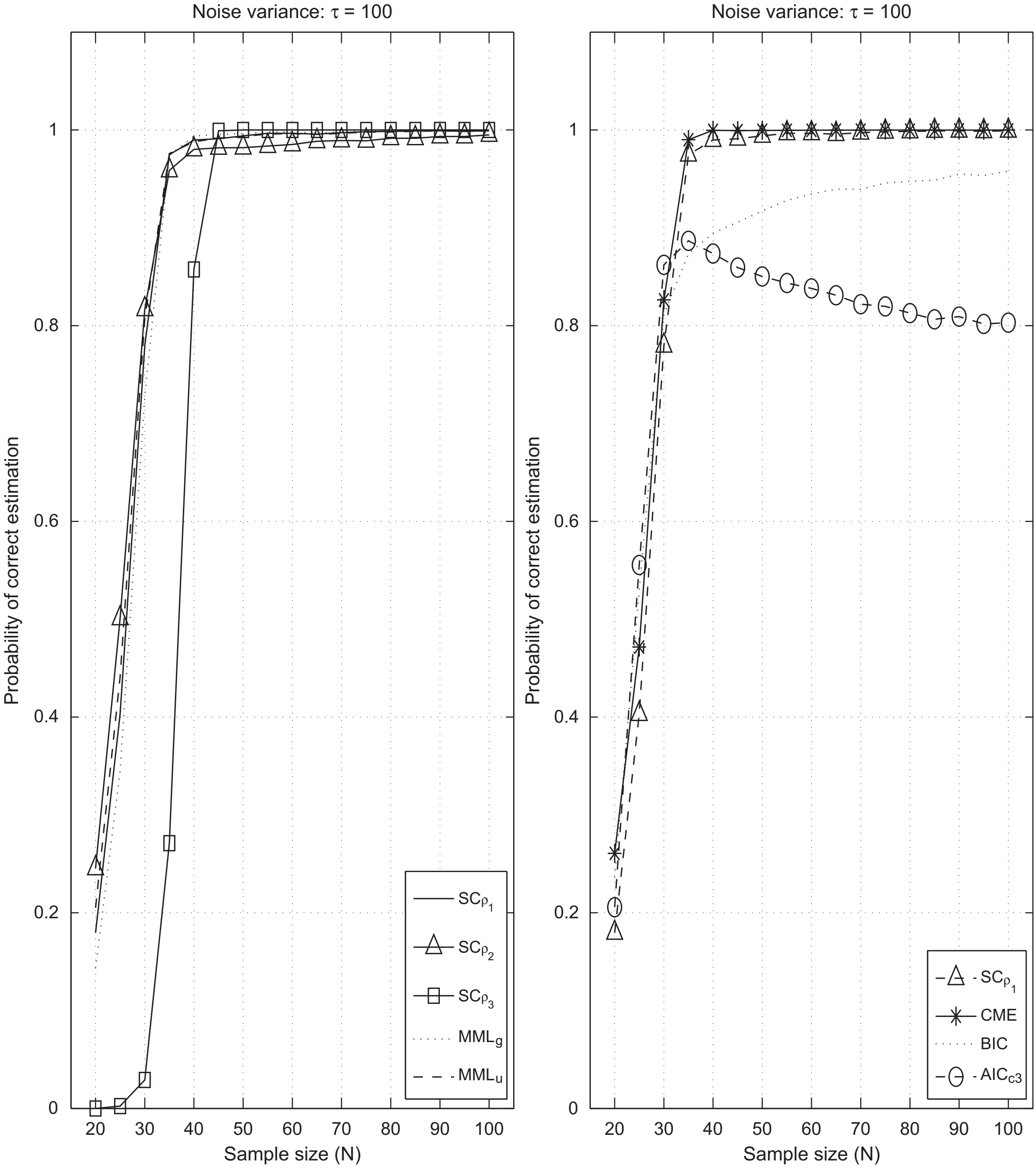

**Fig. 5.** Example 3—the empirical probability of estimating correctly the order of the polynomial model versus the sample size.

observations represent a parabolic signal in noise. In the first scenario, the noise variance is $\tau=100$ and the sample size is varied by choosing $N$ from the set $\{20,25,\ldots,100\}$. Based on $10^4$ trials for each value of $N$, we evaluate the empirical probabilities of selecting the $\gamma_3$-structure, and we plot them in Fig. 5. Then the sample size is kept fixed ($N$=40), and the SNR is varied by modifying the noise variance such that $1/\tau \in \{1/10^3, 2/10^3, \ldots, 10/10^3\}$. The number of runs for each value of $1/\tau$ is $10^4$, and the results are shown in Fig. 6.

Remark in Fig. 5 that the results of $\mathrm{SC}_{\rho_1}$, $\mathrm{SC}_{\rho_2}$, $\mathrm{MML}_\mathrm{g}$ and $\mathrm{MML}_\mathrm{u}$ are very similar for all values of $N$, whereas $\mathrm{SC}_{\rho_3}$ fails to estimate properly the structure when $N \le 40$. Moreover, for $N$=40, $\mathrm{SC}_{\rho_3}$ is inferior to $\mathrm{SC}_{\rho_1}$, $\mathrm{SC}_{\rho_2}$, $\mathrm{MML}_\mathrm{g}$

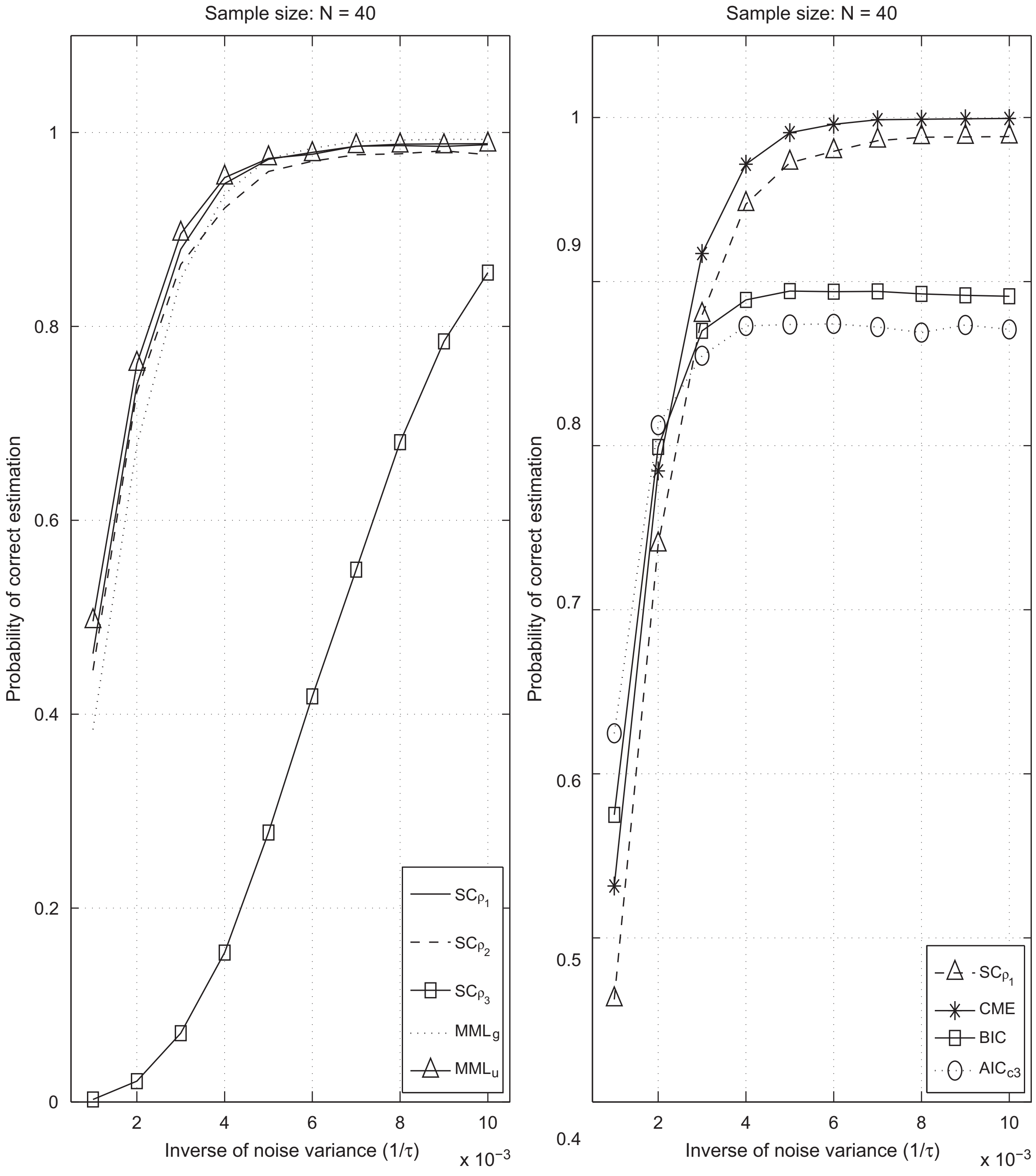

**Fig. 6.** Example 3—the empirical probability of estimating correctly the order of the polynomial model versus the inverse of the noise variance. Note that the range of values being presented along the vertical axes is different for the two plots.

and MML$_u$ for all values of SNR considered in Fig. 6. CME performs extremely well in both figures, and SC$_{\rho_1}$ is almost as good as CME. AIC$_{c3}$ confirms in Fig. 5 what we have already seen in the previous example: it outperforms other criteria when the sample size is small ($N \le 30$), but for large sample size its estimation capabilities are modest. The accuracy of the BIC estimate is better and better when $N$ increases, but even for $N=100$, BIC remains inferior to CME. In Fig. 6, CME outperforms BIC for a large span of SNR values.

The fact that, for the polynomial model, CME is superior to BIC has been already pointed out in [15,16], and it can be understood by resorting to the following asymptotic results from [5]:

$$\mathbf{X}_k^\top \mathbf{X}_k \approx \begin{bmatrix} N & \frac{N^2}{2} & \cdots & \frac{N^k}{k} \\ \frac{N^2}{2} & \frac{N^3}{3} & \cdots & \frac{N^{k+1}}{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{N^k}{k} & \frac{N^{k+1}}{k+1} & \cdots & \frac{N^{2k-1}}{2k-1} \end{bmatrix}, \tag{65}$$

$$|\mathbf{X}_k^\top \mathbf{X}_k| = O(N^{k^2}). \tag{66}$$

Therefore, $\frac{1}{2}\ln|\mathbf{X}_k^\top \mathbf{X}_k|$ which is the penalty term of CME can be written as $[(k^2/2)\ln n + O(1)]$. This shows immediately that $(k/2)\ln n$, the penalty term of BIC, is not the correct one (see [5] for a more detailed discussion). More interestingly, by combining (34) with the approximations from (65)–(66), we have

$$D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1) = \hat{\boldsymbol{\beta}}_k^\top (\mathbf{X}_k^\top \mathbf{X}_k)\hat{\boldsymbol{\beta}}_k = O(N^{2k-1}),$$

$$D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2) = \frac{\|\hat{\boldsymbol{\beta}}_k\|^2}{|\mathbf{X}_k^\top \mathbf{X}_k|^{-1/k}} = O(N^k),$$

$$D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3) = \frac{\hat{\boldsymbol{\beta}}_k^\top (\mathbf{X}_k^\top \mathbf{X}_k)^2 \hat{\boldsymbol{\beta}}_k}{|\mathbf{X}_k^\top \mathbf{X}_k|^{1/k}} = O(N^{3k-2}).$$

According to (33), the penalty term of SC$_{\rho_i}(\mathbf{y};\gamma_k)$ is given by $k\ln D_{\gamma_k}(\mathbf{y};\mathbf{Q}_i)$, where $i \in \{1,2,3\}$. Thus, we can express as follows the penalty terms of the criteria listed below:

$$\begin{aligned} &\mathrm{SC}_{\rho_1}: \ (2k^2-k)\ln n + O(1), \\ &\mathrm{SC}_{\rho_2}: \ k^2 \ln n + O(1), \\ &\mathrm{SC}_{\rho_3}: \ (3k^2-2k)\ln n + O(1). \end{aligned}$$

Recall that the formula in (33) was multiplied by two for writing the equations in a more compact form. Consequently, the above results must be divided by two before comparing them with the penalty term of CME. Note that only SC$_{\rho_2}$ penalizes the complexity of the model as CME does. SC$_{\rho_3}$ is the criterion that deviates the most from the recommended penalty which is $[(k^2/2)\ln n + O(1)]$, and this explains why, in Figs. 5 and 6, the performance of SC$_{\rho_3}$ is modest.

The experimental results obtained for Examples 1–3 lead to some guidance on the application of various model selection criteria to the estimation problems which have been investigated. We summarize the recommendations in Table 1.

*Example* 4 is focused on the predictive capabilities of the model selection criteria which are investigated. Given a data set that contains, for $n$ different instances, the measurements of $m$ input attributes along with the measurements of the response variable, we randomly choose $n_{tr}$ samples to be the training set. Based on the linear regression model, each criterion uses the training set to choose the most relevant input attributes. The model learned by each criterion is applied to the remaining $n-n_{tr}$ samples, which constitute the test set, and the squared prediction error is computed.

The data sets used in our experiments are listed below. For each of them, we indicate the values of $n$ and $m$, as well as the repository where they are publicly available:

1. *Housing* data set: $n=506$, $m=13$, http://archive.ics.uci.edu/ml/datasets/Housing.

**Table 1**
Guidance on the use of the eight criteria for the estimation problems in Examples 1–3: **A**—recommended; B—acceptable; C—unsatisfactory; D—not recommended. For each example, the information about the experimental conditions (sample size and SNR) is provided with the conventions from Figs. 1 to 6.

| Experimental conditions | SC$_{\rho_1}$ | SC$_{\rho_2}$ | SC$_{\rho_3}$ | MML$_g$ | MML$_u$ | CME | BIC | AIC$_{c3}$ |
|---|---|---|---|---|---|---|---|---|
| **Example 1**—Select the variables which are linearly independent | | | | | | | | |
| $n=50$; SNR$=0$ dB | B | C | **A** | B | B | D | B | B |
| **Example 2**—Estimate correctly the number of sinusoids embedded in Gaussian noise | | | | | | | | |
| $N \in [100,150)$; $\tau^{-1}=0.10$ | B | B | B | B | C | D | C | B |
| $N \in [150,200]$; $\tau^{-1}=0.10$ | B | B | B | B | B | D | B | B |
| $N \in (200,300]$; $\tau^{-1}=0.10$ | B | B | B | **A** | **A** | D | **A** | C |
| $N=100$; $\tau^{-1} \in [0.01,0.07]$ | C | C | C | C | C | D | C | C |
| $N=100$; $\tau^{-1} \in (0.07,0.10)$ | C | C | C | C | C | D | C | B |
| $N=100$; $\tau^{-1} \in [0.10,0.12)$ | B | B | B | B | C | D | C | B |
| $N=100$; $\tau^{-1} \in (0.12,0.20]$ | B | B | B | B | B | D | B | B |
| **Example 3**—Estimate correctly the order of a polynomial in Gaussian noise | | | | | | | | |
| $N \in [20,25]$; $\tau^{-1}=0.01$ | C | C | D | C | C | C | C | C |
| $N \in (25,40)$; $\tau^{-1}=0.01$ | B | B | D | B | B | B | B | B |
| $N=40$; $\tau^{-1}=0.01$ | **A** | **A** | C | **A** | **A** | **A** | C | C |
| $N \in (40,100]$; $\tau^{-1}=0.01$ | **A** | **A** | **A** | **A** | **A** | **A** | C | D |
| $N=40$; $\tau^{-1} \in [0.001,0.004)$ | B | B | D | B | B | B | B | B |
| $N=40$; $\tau^{-1} \in [0.004,0.01)$ | **A** | **A** | D | **A** | **A** | **A** | C | C |

2. *Diabetes* data set (standardized): $n=442$, $m=10$, http://www-stat.stanford.edu/~hastie/Papers/LARS.
3. *Concrete* compressive strength data set: $n=1030$, $m=8$, http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength.

At the addresses outlined above, the interested reader can find the data tables, an accurate description of their content, along with references to previous works where they have been utilized. For instance, all three data sets have been also used for the experimental part of [28]. We note that, in [28], the Housing data set has been altered as follows: the measurements corresponding to the attribute CHAS (Charles River dummy variable) have been removed, the values of the attribute NOX (nitric oxides concentration—parts per 10 million) have been multiplied by 100, and the $\mathbf{y}$-vector for the response variable has been transformed such that to have zero-mean. Similarly, the vector for the response variable within Concrete compressive strength data set has been modified to have zero-mean. Because we want our settings to be like in [28], we apply the same changes. Moreover, the model selection during the training step is slightly different than how it was performed in Examples 1–3:

- One modification is that we select the best model among all $\gamma$-structures which are subsets of $\{1,\ldots,m\}$, including the case $\gamma=\emptyset$. Therefore, the tested models are not nested, and we cannot any longer apply the recursive least-squares algorithm [14, p. 237] to estimate the linear parameters, as it was done in Examples 1–3. Like in [28], we use the Moore–Penrose pseudoinverse.
- Another modification is that we do *not* neglect the term $L(\gamma)$ which quantifies the complexity of the structure. To be in line with [28], we do not apply the formula from [26], but the following one:

$$L'(\gamma)=\ln\binom{m}{k}+\ln(m+1),$$

where $k$ denotes the cardinality of $\gamma$. Obviously, $2L'(\gamma)$ is added to $\mathrm{SC}_{\rho_1}$, $\mathrm{SC}_{\rho_2}$, $\mathrm{SC}_{\rho_3}$, and $L'(\gamma)$ is added to the other criteria.

The predictive accuracy is evaluated for five different values of $n_{\mathrm{tr}}$, and the results are shown in Table 2. Note that each entry within Table 2 is calculated as an average of the prediction errors obtained from $10^3$ random partitions of the data sets into training/test subsets. The results for $\mathrm{MML_g}$ and $\mathrm{MML_u}$ are identical with those from [28]. Because in [28], it was not used the Stirling approximation (26) when evaluating $\mathrm{SC}_{\rho_1}$, for this criterion, there exist small differences between the results from Table 2 and the results reported in [28].

Based on the empirical evidence, it is not possible to decide that one particular criterion has stronger prediction capabilities than the others. It is interesting to remark in Table 2 that it does not exist any combination of experimental settings for which BIC yields the smallest prediction error. The same is true for $\mathrm{SC}_{\rho_1}$. Overall, $\mathrm{SC}_{\rho_2}$ is slightly superior to $\mathrm{SC}_{\rho_1}$. CME performs surprisingly well for the Diabetes data set, but for the other two data sets, its results are moderate.

## 5. Conclusions

In the case of the Gaussian linear regression, the parametric complexity is not finite and the only possibility for obtaining NML-based selection rules is to constrain the data space. Even if this was recognized long time ago, the solutions proposed so far are only punctual results which treat some particular constraints. In this paper, we have introduced a general methodology for addressing the problem. Based on the new findings, we demonstrated how the rhomboidal constraint yields a new NML-based formula. Additionally, we used the ellipsoidal constraint to re-derive three criteria that have been introduced in the previous literature: $\mathrm{SC}_{\rho_1}$ [25] and $\mathrm{SC}_{\rho_2}$ and $\mathrm{SC}_{\rho_3}$ [18]. They have been compared against BIC [29], $\mathrm{AIC_{c3}}$ [32], CME [15] and $\mathrm{MML_g}$ and $\mathrm{MML_u}$ [28].

The theoretical analysis and the Monte Carlo simulations led to the following outcomes: (a) $\mathrm{SC}_{\rho_3}$ has the strongest tendency to select the variables which are linearly independent; (b) $\mathrm{SC}_{\rho_1}$, $\mathrm{SC}_{\rho_2}$ and $\mathrm{SC}_{\rho_3}$ reduce to one single criterion when the regression matrix is orthonormal; (c) $\mathrm{MML_g}$ and

**Table 2**
Example 4—squared prediction errors obtained for real life measurements. For all data sets, it is written in bold the best result for each $n_{\mathrm{tr}}$.

| Data set | $n_{\mathrm{tr}}$ | $\mathrm{SC}_{\rho_1}$ | $\mathrm{SC}_{\rho_2}$ | $\mathrm{SC}_{\rho_3}$ | $\mathrm{MML_g}$ | $\mathrm{MML_u}$ | CME | BIC | $\mathrm{AIC_{c3}}$ |
|---|---|---|---|---|---|---|---|---|---|
| Housing | 25 | 69.976 | **52.529** | 53.249 | 61.922 | 71.509 | 85.282 | 70.326 | 59.463 |
| | 50 | 36.933 | **35.265** | 37.268 | 36.340 | 36.635 | 36.147 | 36.511 | 36.577 |
| | 100 | 29.323 | 30.210 | 30.523 | 29.624 | 29.383 | 29.079 | 29.516 | **28.343** |
| | 200 | 26.023 | 27.711 | 27.657 | 26.424 | 26.162 | 26.897 | 26.535 | **25.271** |
| | 400 | 24.315 | 25.998 | 26.225 | 24.304 | **24.299** | 24.645 | 24.365 | 24.321 |
| Diabetes | 25 | 4824.3 | **4362.9** | 4553.0 | 4445.0 | 4819.2 | 5386.5 | 4647.5 | 4506.3 |
| | 50 | 3855.3 | **3645.3** | 3902.0 | 3851.2 | 3843.8 | 3722.5 | 3819.5 | 3743.1 |
| | 100 | 3355.2 | 3259.9 | 3410.1 | 3385.3 | 3364.2 | **3237.2** | 3368.4 | 3301.5 |
| | 200 | 3165.9 | 3099.7 | 3210.7 | 3199.6 | 3173.3 | **3069.5** | 3195.4 | 3073.4 |
| | 400 | 3046.9 | 3060.5 | 3053.4 | 3052.8 | 3052.7 | **3026.9** | 3055.7 | **3026.9** |
| Concrete | 25 | 225.18 | 257.71 | 245.86 | **221.2** | 227.41 | 279.27 | 235.07 | 245.40 |
| | 50 | 148.67 | 148.57 | **147.11** | 147.46 | 149.25 | 162.36 | 150.06 | 148.78 |
| | 100 | 123.82 | **121.56** | 121.59 | 122.90 | 123.65 | 124.00 | 123.29 | 124.11 |
| | 200 | 114.56 | **113.89** | 114.05 | 114.37 | 114.50 | 114.17 | 114.31 | 114.89 |
| | 400 | 111.67 | **111.12** | 111.46 | 111.59 | 111.64 | 111.22 | 111.56 | 111.70 |

MML$_u$ perform similarly with SC$_{\rho_1}$ and they are superior to SC$_{\rho_1}$ for some particular experimental settings; (d) AIC$_{c3}$ is very good when the sample size is small, but it has modest results when the sample size is large; (e) BIC has a behavior which is opposite to that of AIC$_{c3}$, and the performance of SC$_{\rho_1}$ is an excellent compromise between BIC and AIC$_{c3}$; (f) CME poses troubles for some models, but in the case of the polynomial model, it is ranked the first for a large range of sample sizes; (g) SC$_{\rho_1}$, SC$_{\rho_2}$, MML$_g$ and MML$_u$ are nearly as good as CME for the polynomial model, while SC$_{\rho_3}$ has difficulties in this case.

## Acknowledgments

The work of C.D. Giurcăneanu and S.A. Razavi was supported by the Academy of Finland, Project Nos. 113572, 118355, 134767 and 213462. The work of A. Liski was supported by the Academy of Finland under Project No. 213462. S.A. Razavi is grateful to Dr. Enes Makalic for providing the Matlab code used in the experimental part of [28].

## Appendix A. Evaluation of the normalized maximum likelihood

The techniques that we apply in this section are very similar with those used in [17,18,25–27].

*Computation of* $C_\rho(R,\tau_0)$. First we note that the estimated parameter vector $[\hat{\boldsymbol{\beta}}^\top\ \hat{\tau}]^\top$ is a sufficient statistic, and the density $f(\mathbf{y};\boldsymbol{\beta},\tau)$ can be factored as follows:

$$f(\mathbf{y};\boldsymbol{\beta},\tau)=f(\mathbf{y}|\hat{\boldsymbol{\beta}},\hat{\tau})g(\boldsymbol{\beta},\tau;\hat{\boldsymbol{\beta}},\hat{\tau}), \tag{A.1}$$

where $f(\mathbf{y}|\hat{\boldsymbol{\beta}},\hat{\tau})$ does not depend on the unknowns $\boldsymbol{\beta}$ and $\tau$. According to [31, Theorem 3.5], the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}$ are statistically independent, and we have

$$g(\boldsymbol{\beta},\tau;\hat{\boldsymbol{\beta}},\hat{\tau})=g_1(\hat{\boldsymbol{\beta}};\boldsymbol{\beta},\tau)g_2(\hat{\tau};\tau),$$

$$g_1(\hat{\boldsymbol{\beta}};\boldsymbol{\beta},\tau)=\frac{|\mathbf{X}^\top\mathbf{X}|^{1/2}}{(2\pi\tau)^{k/2}}\exp\left(-\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2}{2\tau}\right),$$

$$g_2(\hat{\tau};\tau)=\frac{n^{(n-k)/2}}{\Gamma\left(\frac{n-k}{2}\right)2^{(n-k)/2}}\left(\frac{\hat{\tau}}{\tau}\right)^{(n-k)/2}\frac{1}{\hat{\tau}}\exp\left(-\frac{n\hat{\tau}}{2\tau}\right).$$

By employing (15), we obtain

$$g(\hat{\boldsymbol{\beta}},\hat{\tau};\hat{\boldsymbol{\beta}},\hat{\tau})=A_{n,k}\hat{\tau}^{-k/2-1}. \tag{A.2}$$

Then we define $\mathcal{P}(R,\tau_0)=\{[\hat{\boldsymbol{\beta}}^\top\ \hat{\tau}]^\top:\rho(\hat{\boldsymbol{\beta}})\le R,\hat{\tau}\ge\tau_0\}$ and $\mathcal{Y}(\hat{\boldsymbol{\beta}},\hat{\tau})=\{\mathbf{y}:\hat{\boldsymbol{\beta}}(\mathbf{y})=\hat{\boldsymbol{\beta}},\hat{\tau}(\mathbf{y})=\hat{\tau}\}$. After these preparations, we evaluate the integral in (12):

$$C_\rho(R,\tau_0)=\int_{\mathcal{P}(R,\tau_0)}\left[\int_{\mathcal{Y}(\hat{\boldsymbol{\beta}},\hat{\tau})}f(\mathbf{y}|\hat{\boldsymbol{\beta}},\hat{\tau})\,\mathrm{d}\mathbf{y}\right]g(\hat{\boldsymbol{\beta}},\hat{\tau};\hat{\boldsymbol{\beta}},\hat{\tau})\,\mathrm{d}\hat{\boldsymbol{\beta}}\,\mathrm{d}\hat{\tau} \tag{A.3}$$

$$=A_{n,k}\int_{\tau_0}^{\infty}\hat{\tau}^{-k/2-1}\,\mathrm{d}\hat{\tau}\int_{\mathcal{B}(R)}\mathrm{d}\hat{\boldsymbol{\beta}}=(2A_{n,k}/k)\tau_0^{-k/2}V_\rho(R). \tag{A.4}$$

Remark in (A.3) that the inner integral gives unity [26]. The use of (A.2) and some simple manipulations yield (A.4). Additionally, (10) and (A.4) lead to (14).

*Computation of* $\overline{C}_\rho(R_1,R_2,\tau_1,\tau_2)$. For evaluating the normalizing constant in (18), we define $\mathcal{P}(R_1,R_2,\tau_1,\tau_2)=\{[\hat{\boldsymbol{\beta}}^\top\ \hat{\tau}]^\top:R_1\le\rho(\hat{\boldsymbol{\beta}})\le R_2,\tau_1\le\hat{\tau}\le\tau_2\}$ and $\mathcal{B}(R_1,R_2)=\{\hat{\boldsymbol{\beta}}:R_1\le\rho(\hat{\boldsymbol{\beta}})\le R_2\}$. So,

$$\overline{C}_\rho(R_1,R_2,\tau_1,\tau_2)=\int_{\mathcal{Y}(R_1,R_2,\tau_1,\tau_2)}\hat{f}(\mathbf{y};\tilde{R},\tilde{\tau}_0)\,\mathrm{d}\mathbf{y}$$

$$=\int_{\mathcal{Y}(R_1,R_2,\tau_1,\tau_2)}\frac{f(\mathbf{y}|\hat{\boldsymbol{\beta}},\hat{\tau})g(\hat{\boldsymbol{\beta}},\hat{\tau};\hat{\boldsymbol{\beta}},\hat{\tau})}{C_\rho(\tilde{R},\tilde{\tau}_0)}\,\mathrm{d}\mathbf{y} \tag{A.5}$$

$$=\int_{\mathcal{P}(R_1,R_2,\tau_1,\tau_2)}\frac{g(\hat{\boldsymbol{\beta}},\hat{\tau};\hat{\boldsymbol{\beta}},\hat{\tau})}{C_\rho(\tilde{R},\tilde{\tau}_0)}\left[\int_{\mathcal{Y}(\hat{\boldsymbol{\beta}},\hat{\tau})}f(\mathbf{y}|\hat{\boldsymbol{\beta}},\hat{\tau})\,\mathrm{d}\mathbf{y}\right]\mathrm{d}\hat{\boldsymbol{\beta}}\,\mathrm{d}\hat{\tau} \tag{A.6}$$

$$=\int_{\mathcal{P}(R_1,R_2,\tau_1,\tau_2)}\frac{A_{n,k}\hat{\tau}^{-k/2-1}}{A_{n,k}(2/k)\hat{\tau}^{-k/2}V_\rho(\tilde{R})}\,\mathrm{d}\hat{\boldsymbol{\beta}}\,\mathrm{d}\hat{\tau} \tag{A.7}$$

$$=\frac{k}{2}\int_{\tau_1}^{\tau_2}\frac{1}{\hat{\tau}}\,\mathrm{d}\hat{\tau}\int_{\mathcal{B}(R_1,R_2)}\eta^{-1}[\rho(\hat{\boldsymbol{\beta}})]^{-\zeta k}\,\mathrm{d}\hat{\boldsymbol{\beta}} \tag{A.8}$$

$$=\frac{k}{2}\ln\frac{\tau_2}{\tau_1}\int_{R_1}^{R_2}\frac{(\eta\zeta k)R^{\zeta k-1}}{\eta R^{\zeta k}}\,\mathrm{d}R$$
$$=\frac{\zeta k^2}{2}\ln\frac{\tau_2}{\tau_1}\ln\frac{R_2}{R_1}.$$

Note that in (A.5) we use again the factorization from (A.1). Similarly with (A.3), the inner integral in (A.6) gives unity. The identity in (A.7) is derived straightforwardly from (A.2), (A.4) and (A.6). For the calculation of the second integral in (A.8), we apply the same technique as in [25,26] and, based on (10), we take the element of volume to be $\mathrm{d}V_\rho=\eta\zeta kR^{\zeta k-1}\,\mathrm{d}R$. After some simple algebra, we get the result in (19).

*Evaluation of the approximate formula* (23). Note that the approximation from (23) can be applied for a much more general class of models, and not only for the model in (2). The proof given in [24] treats the general case and is based on sophisticated mathematical derivations. However, it was already pointed out in [26, Section 5.2.2] that the proof can be simplified if the analyzed model satisfies a particular condition. With our notations, the condition is as follows:

$$\lim_{n\to\infty}\frac{g(\hat{\boldsymbol{\beta}},\hat{\tau};\hat{\boldsymbol{\beta}},\hat{\tau})}{[n/(2\pi)]^{(k+1)/2}|\mathbf{J}_n(\hat{\boldsymbol{\beta}},\hat{\tau})|^{1/2}}=1. \tag{A.9}$$

Observe that Eq. (25) leads to

$$\left(\frac{n}{2\pi}\right)^{(k+1)/2}|\mathbf{J}_n(\hat{\boldsymbol{\beta}},\hat{\tau})|^{1/2}=\check{A}_{n,k}\hat{\tau}^{-k/2-1},$$

$$\check{A}_{n,k}=\frac{|\mathbf{X}^\top\mathbf{X}|^{1/2}\sqrt{n}}{(2\pi)^{(k+1)/2}\sqrt{2}}.$$

By using (15) and (A.2), we get

$$\lim_{n\to\infty}\frac{g(\hat{\boldsymbol{\beta}},\hat{\tau};\hat{\boldsymbol{\beta}},\hat{\tau})}{[n/(2\pi)]^{(k+1)/2}|\mathbf{J}_n(\hat{\boldsymbol{\beta}},\hat{\tau})|^{1/2}}=\lim_{n\to\infty}\frac{A_{n,k}}{\check{A}_{n,k}}$$

$$= \lim_{n\to\infty} \frac{n^{(n-k-1)/2}\sqrt{2\pi}}{2^{(n-k-1)/2}\exp\left(\frac{n}{2}\right)\Gamma\left(\frac{n-k}{2}\right)}$$

$$= \lim_{n\to\infty} \frac{\left(1-\frac{k}{n}\right)^{(k+1)/2}}{\left(1-\frac{k/2}{n/2}\right)^{n/2}\exp\left(\frac{k}{2}\right)} = 1. \tag{A.10}$$

The identity in (A.10) was obtained by taking $z=(n-k)/2$ in the well-known expression of the Gamma function:

$$\Gamma(z) = z^{z-1/2}\exp(-z)\exp[\mu(z)]\sqrt{2\pi}, \tag{A.11}$$

where $\mu(z)=\overline{\mu}/(12z)$ and $\overline{\mu}\in(0,1)$. Remark that the Stirling approximation in (26) is a straightforward consequence of (A.11).

Our approach is slightly different than the one from [26, Section 5.2.2] where the condition (A.9) was employed to prove (23). More precisely, we consider the following asymptotic approximation:

$$g(\hat{\boldsymbol\beta},\hat{\tau};\hat{\boldsymbol\beta},\hat{\tau}) \approx \left(\frac{n}{2\pi}\right)^{(k+1)/2} |\mathbf{J}_\infty(\hat{\boldsymbol\beta},\hat{\tau})|^{1/2} = \check{A}_{\infty,k}\hat{\tau}^{-k/2-1}, \tag{A.12}$$

where

$$\check{A}_{\infty,k} = \left(\frac{n}{2\pi}\right)^{(k+1)/2}\left(\frac{|\mathbf{G}_\infty|}{2}\right)^{1/2},$$

$$\mathbf{G}_\infty = \lim_{n\to\infty}\mathbf{G}_n, \tag{A.13}$$

$$\mathbf{G}_n = \frac{\mathbf{X}^\top\mathbf{X}}{n}. \tag{A.14}$$

Because we want to apply the same techniques like in the evaluation of $\hat{f}_\rho(\mathbf{y})$, we use in (A.3) the approximation from (A.12), which leads to

$$C_\rho^{\mathrm{FIM}}(R,\tau_0) = (2\check{A}_{\infty,k}/k)\tau_0^{-k/2}V_\rho(R). \tag{A.15}$$

It is important to remark that the expression of $\overline{C}_\rho(R_1,R_2,\tau_1,\tau_2)$ remains unchanged when (A.5) is modified as follows: (i) $g(\hat{\boldsymbol\beta},\hat{\tau};\hat{\boldsymbol\beta},\hat{\tau})$ is replaced by the approximation given in (A.12); (ii) $C_\rho(\tilde{R},\tilde{\tau}_0)$ is replaced by $C_\rho^{\mathrm{FIM}}(\tilde{R},\tilde{\tau}_0)$. Consequently, the approximate formula of SC is

$$-\ln\hat{f}_\rho^{\mathrm{FIM}}(\mathbf{y}) = -\ln f(\mathbf{y};\hat{\boldsymbol\beta},\hat{\tau}) + \ln C_\rho^{\mathrm{FIM}}(\rho(\hat{\boldsymbol\beta}),\hat{\tau}) + \ln\overline{C}_\rho(R_1,R_2,\tau_1,\tau_2).$$

Furthermore, we compare this result with the one from (20):

$$-\ln\frac{\hat{f}_\rho(\mathbf{y})}{\hat{f}_\rho^{\mathrm{FIM}}(\mathbf{y})} = \ln\frac{C_\rho(\rho(\hat{\boldsymbol\beta}),\hat{\tau})}{C_\rho^{\mathrm{FIM}}(\rho(\hat{\boldsymbol\beta}),\hat{\tau})} = \ln\frac{A_{n,k}}{\check{A}_{\infty,k}}$$

$$= \frac{1}{2}\ln\frac{|\mathbf{G}_n|}{|\mathbf{G}_\infty|} + \frac{n-k-1}{2}\ln\frac{n}{2} - \ln\Gamma\left(\frac{n-k}{2}\right) + \frac{1}{2}\ln(2\pi) - \frac{n}{2} \tag{A.16}$$

$$\approx \frac{1}{2}\ln\frac{|\mathbf{G}_n|}{|\mathbf{G}_\infty|} - \frac{n-k-1}{2}\ln\frac{n-k}{n} - \frac{k}{2}. \tag{A.17}$$

Eq. (A.16) shows clearly that the difference $[-\ln\hat{f}_\rho(\mathbf{y})] - [-\ln\hat{f}_\rho^{\mathrm{FIM}}(\mathbf{y})]$ does not depend on the constraint $\rho(\cdot)$ which is used for computing the integral. Moreover, based on (A.10), (A.13), (A.14), (A.17), it is easy to conclude that $\hat{f}_\rho(\mathbf{y})$ and $\hat{f}_\rho^{\mathrm{FIM}}(\mathbf{y})$ are the same when $n$ is large. Note that the derivation of (A.17) involves the Stirling approximation (26).

## Appendix B. Proofs of the main results within Section 3.2

**Proof of Proposition 3.1.**

(a) We consider the singular value decomposition (SVD) of the matrix $\mathbf{X}_\gamma$. Let $\mathbf{X}_\gamma = [\mathbf{U}\ \mathbf{U}_0][\boldsymbol\Lambda^\top\ \mathbf{0}^\top]^\top\mathbf{V}^\top$, where the matrix $[\mathbf{U}\ \mathbf{U}_0]$ has orthonormal columns, $\mathbf{U}\in\mathbb{R}^{n\times k}$ and $\mathbf{U}_0\in\mathbb{R}^{n\times(n-k)}$. The diagonal matrix $\boldsymbol\Lambda\in\mathbb{R}^{k\times k}$ is non-singular, and $\mathbf{V}\in\mathbb{R}^{k\times k}$ is such that $\mathbf{V}^{-1}=\mathbf{V}^\top$. For $i\in\{1,2,3\}$, we have $\mathbf{Q}_i = \mathbf{V}\mathbf{L}_i^2\mathbf{V}^\top$ and

$$D_\gamma(\mathbf{y};\mathbf{Q}_i) = \mathbf{y}^\top\mathbf{U}\frac{\mathbf{L}_i^2}{|\mathbf{L}_i^2|^{1/k}}\mathbf{M}^{-1}\mathbf{U}^\top\mathbf{y},$$

where $\mathbf{L}_1=\boldsymbol\Lambda$, $\mathbf{L}_2=\mathbf{I}$, $\mathbf{L}_3=\boldsymbol\Lambda^2$ and $\mathbf{M}=\boldsymbol\Lambda^2/|\boldsymbol\Lambda^2|^{1/k}$. The equalities in (35) can be re-written as

$$\|\mathbf{U}^\top\mathbf{y}\|^2 = \|\mathbf{M}^{-1/2}\mathbf{U}^\top\mathbf{y}\|^2 = \|\mathbf{M}^{1/2}\mathbf{U}^\top\mathbf{y}\|^2,$$

and they are satisfied for all $\mathbf{y}\in\mathbb{R}^n\backslash\{\mathbf{0}\}$ if and only if $\mathbf{M}=\mathbf{I}$. This is equivalent with the fact that $\boldsymbol\Lambda^2$ has one eigenvalue with multiplicity $k$. We denote $q$ this eigenvalue, and the condition in (36) is immediately obtained.

(b) We use the notations introduced in the proof of the point (a), and we focus on the properties of the matrix $\mathbf{M}$. Observe that the diagonal entries of $\mathbf{M}$ are strictly positive, and their product is equal to one. If $\mathbf{M}\neq\mathbf{I}$, then some of the eigenvalues of $\mathbf{M}^{-1}$ are larger than one, while the others are smaller than one. Therefore, the matrix $\mathbf{I}-\mathbf{M}^{-1}$ has both positive and negative eigenvalues. This observation together with the identity $D_\gamma(\mathbf{y};\mathbf{Q}_1)-D_\gamma(\mathbf{y};\mathbf{Q}_2) = \mathbf{y}^\top\mathbf{U}(\mathbf{I}-\mathbf{M}^{-1})\mathbf{U}^\top\mathbf{y}$ show that, depending on $\mathbf{y}$, the difference $D_\gamma(\mathbf{y};\mathbf{Q}_1)-D_\gamma(\mathbf{y};\mathbf{Q}_2)$ can be either positive or negative. The proof is similar for $D_\gamma(\mathbf{y};\mathbf{Q}_2)-D_\gamma(\mathbf{y};\mathbf{Q}_3)$ and $D_\gamma(\mathbf{y};\mathbf{Q}_1)-D_\gamma(\mathbf{y};\mathbf{Q}_3)$.

(c) It is easy to verify that $D_\gamma(\mathbf{y};\mathbf{Q}_2)\times D_\gamma(\mathbf{y};\mathbf{Q}_3) = \|\hat{\boldsymbol\beta}_\gamma\|^2\times\|\mathbf{X}_\gamma^\top\mathbf{y}\|^2$ and $D_\gamma(\mathbf{y};\mathbf{Q}_1)^2 = [\hat{\boldsymbol\beta}_\gamma^\top(\mathbf{X}_\gamma^\top\mathbf{y})]^2$. The Cauchy–Schwarz inequality [30, p. 258] written for the vectors $\hat{\boldsymbol\beta}_\gamma$ and $\mathbf{X}_\gamma^\top\mathbf{y}$ leads to $D_\gamma(\mathbf{y};\mathbf{Q}_2)\times D_\gamma(\mathbf{y};\mathbf{Q}_3)\geq D_\gamma(\mathbf{y};\mathbf{Q}_1)^2$, which proves the inequality in (37). □

**Proof of Proposition 3.2.** The main idea is to write the expressions of $D_\gamma(\mathbf{y};\mathbf{Q}_i)$, $i\in\{1,2,3\}$, in a form which allows us to see immediately if, for $\alpha\to0$, the result is finite or not. We introduce the following supplementary notations: $\mathbf{g}=\mathbf{P}_{k\text{-}1}^\perp\mathbf{x}_k$, $\mathbf{P}_k=\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top$, $\mathbf{P}_{\mathbf{x}_k}=(\mathbf{x}_k\mathbf{x}_k^\top)/\|\mathbf{x}_k\|^2$ and $\mathbf{P}_{\mathbf{x}_k}^\perp=\mathbf{I}-\mathbf{P}_{\mathbf{x}_k}$. The symbol $^\#$ is used for the Moore–Penrose pseudoinverse.

(a) Some simple manipulations combined with the identity from [14, Eq. (8.34)] lead to

$$D_\gamma(\mathbf{y};\mathbf{Q}_1) = \|\mathbf{P}_k\mathbf{y}\|^2 = \left\|\left(\mathbf{P}_{k\text{-}1} + \frac{\mathbf{g}}{\|\mathbf{g}\|}\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}\right)\mathbf{y}\right\|^2$$

$$= \|(\mathbf{P}_{k\text{-}1} + \mathbf{w}\mathbf{w}^\top)\mathbf{y}\|^2. \tag{B.1}$$

(b) To compute $D_\gamma(\mathbf{y};\mathbf{Q}_2)$, we use the formula [20, Eq. (2.17)]:

$$(\mathbf{X}_k^\top\mathbf{X}_k)^{-1} = \begin{bmatrix} (\mathbf{X}_{k\text{-}1}^\top\mathbf{P}_{\mathbf{x}_k}^\perp\mathbf{X}_{k\text{-}1})^{-1} & \mathbf{F} \\ \mathbf{F}^\top & (\mathbf{x}_k^\top\mathbf{P}_{k\text{-}1}^\perp\mathbf{x}_k)^{-1}\end{bmatrix}, \tag{B.2}$$

where $\mathbf{F}=-\mathbf{X}_{k\text{-}1}^\#\mathbf{x}_k(\mathbf{x}_k^\top\mathbf{P}_{k\text{-}1}^\perp\mathbf{x}_k)^{-1}$. Simple calculations produce the following outcome:

$$(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top = \begin{bmatrix} \left(\mathbf{P}_{\mathbf{x}_k}^\perp\mathbf{X}_{k\text{-}1}\right)^\# \\ \frac{1}{\|\mathbf{g}\|}\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}\end{bmatrix}. \tag{B.3}$$

Then we employ the identity from [2, Eq. (17)] to get

$$\begin{aligned}(\mathbf{P}_{\mathbf{x}_k}^\perp\mathbf{X}_{k\text{-}1})^\# &= (\mathbf{X}_{k\text{-}1}^\top\mathbf{P}_{\mathbf{x}_k}^\perp\mathbf{X}_{k\text{-}1})^{-1}\mathbf{X}_{k\text{-}1}^\top\mathbf{P}_{\mathbf{x}_k}^\perp \\ &= (\mathbf{X}_{k\text{-}1}^\#\mathbf{X}_{k\text{-}1})(\mathbf{X}_{k\text{-}1}^\top\mathbf{P}_{\mathbf{x}_k}^\perp\mathbf{X}_{k\text{-}1})^{-1}\mathbf{X}_{k\text{-}1}^\top\mathbf{P}_{\mathbf{x}_k}^\perp \\ &= \mathbf{X}_{k\text{-}1}^\#\mathbf{P}_{k\text{-}1}[\mathbf{I}-\mathbf{x}_k(\mathbf{x}_k^\top\mathbf{P}_{k\text{-}1}^\perp\mathbf{x}_k)^{-1}\mathbf{x}_k^\top\mathbf{P}_{k\text{-}1}^\perp] \\ &= \frac{\mathbf{X}_{k\text{-}1}^\#}{\|\mathbf{g}\|}\left(\|\mathbf{g}\|\mathbf{I}-\mathbf{x}_k\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}\right).\end{aligned}$$

The result above together with (B.3) show that

$$\begin{aligned}\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-2}\mathbf{X}_k^\top = \frac{1}{\|\mathbf{g}\|^2}&\left(\|\mathbf{g}\|\mathbf{I}-\mathbf{x}_k\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}\right)^\top(\mathbf{X}_{k\text{-}1}^\#)^\top \\ &\times\mathbf{X}_{k\text{-}1}^\#\left(\|\mathbf{g}\|\mathbf{I}-\mathbf{x}_k\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}\right)+\frac{1}{\|\mathbf{g}\|^2}\frac{\mathbf{g}}{\|\mathbf{g}\|}\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}.\end{aligned} \tag{B.4}$$

Additionally, it is known that [16]

$$|\mathbf{X}_k^\top\mathbf{X}_k| = \|\mathbf{g}\|^2|\mathbf{X}_{k\text{-}1}^\top\mathbf{X}_{k\text{-}1}|. \tag{B.5}$$

So,

$$\begin{aligned}D_\gamma(\mathbf{y};\mathbf{Q}_2) = \frac{|\mathbf{X}_{k\text{-}1}^\top\mathbf{X}_{k\text{-}1}|^{1/k}}{\|\mathbf{g}\|^{2(1-1/k)}}&\left[\left\|\mathbf{X}_{k\text{-}1}^\#\left(\|\mathbf{g}\|\mathbf{I}-\mathbf{x}_k\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}\right)\mathbf{y}\right\|^2\right. \\ &\left.+\left(\frac{\mathbf{g}^\top}{\|\mathbf{g}\|}\mathbf{y}\right)^2\right] = \frac{|\mathbf{X}_{k\text{-}1}^\top\mathbf{X}_{k\text{-}1}|^{1/k}}{[\sin(\alpha)\|\mathbf{x}_k\|]^{2(1-1/k)}} \\ &\times[\|\mathbf{X}_{k\text{-}1}^\#(\sin(\alpha)\|\mathbf{x}_k\|\mathbf{I}-\mathbf{x}_k\mathbf{w}^\top)\mathbf{y}\|^2+(\mathbf{w}^\top\mathbf{y})^2].\end{aligned} \tag{B.6}$$

(c) It is obvious that $D_\gamma(\mathbf{y};\mathbf{Q}_3)=\|\mathbf{X}_k^\top\mathbf{y}\|^2/|\mathbf{X}_k^\top\mathbf{X}_k|^{1/k}$. Then we apply (B.5) to get

$$\begin{aligned}D_\gamma(\mathbf{y};\mathbf{Q}_3) &= \frac{1}{\|\mathbf{g}\|^{2/k}}\frac{\|\mathbf{X}_k^\top\mathbf{y}\|^2}{|\mathbf{X}_{k\text{-}1}^\top\mathbf{X}_{k\text{-}1}|^{1/k}} \\ &= \frac{1}{[\sin(\alpha)\|\mathbf{x}_k\|]^{2/k}}\frac{\|\mathbf{X}_k^\top\mathbf{y}\|^2}{|\mathbf{X}_{k\text{-}1}^\top\mathbf{X}_{k\text{-}1}|^{1/k}}.\end{aligned} \tag{B.7}$$

Proposition 3.2 is a straightforward consequence of (B.1), (B.6) and (B.7). □

**Proof of Proposition 3.3.** The equality in (41) is readily obtained from (34) and (39). We also have from (34) that

$$D_\gamma(\mathbf{y};\mathbf{Q}_2) = \|\hat{\boldsymbol{\beta}}_\gamma\|^2\times|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|^{1/k}. \tag{B.8}$$

Let $\hat{\beta}_k$ be the last entry of the vector $\hat{\boldsymbol{\beta}}_\gamma$. With the notations from the proof of Proposition 3.2, we have

$$\hat{\boldsymbol{\beta}}_k^2 = \left(\frac{\mathbf{g}^\top\mathbf{y}}{\|\mathbf{g}\|^2}\right)^2 \tag{B.9}$$

$$= \frac{\mathbf{y}^\top[(\mathbf{g}\mathbf{g}^\top)/\|\mathbf{g}\|^2]\mathbf{y}}{\|\mathbf{g}\|^2} = \frac{\mathbf{y}^\top(\mathbf{P}_k-\mathbf{P}_{k\text{-}1})\mathbf{y}}{\|\mathbf{P}_k^\perp\mathbf{x}_k\|^2} \tag{B.10}$$

$$= \frac{\|\mathbf{P}_k\mathbf{y}\|^2-\|\mathbf{P}_{k\text{-}1}\mathbf{y}\|^2}{1-\|\mathbf{P}_{k\text{-}1}\mathbf{x}_k\|^2} = \frac{R^2_{\mathbf{y}\cdot\mathbf{X}_\gamma}-R^2_{\mathbf{y}\cdot\mathbf{X}_{\gamma\setminus\{k\}}}}{1-R^2_{k\cdot 1,\ldots,(k-1)}}.$$

Note that (B.9) is obtained from (B.3), and (B.10) is based on (B.1). The result can be extended to all entries of $\hat{\boldsymbol{\beta}}_\gamma$, and we get

$$\|\hat{\boldsymbol{\beta}}_\gamma\|^2 = \sum_{i=1}^{k}\frac{R^2_{\mathbf{y}\cdot\mathbf{X}_\gamma}-R^2_{\mathbf{y}\cdot\mathbf{X}_{\gamma\setminus\{i\}}}}{1-R^2_{\varsigma(k)\cdot\varsigma(1),\ldots,\varsigma(k-1)}}, \tag{B.11}$$

where $\varsigma(\cdot)$ is defined in (44). Next we use recursively the identity from (B.5) to obtain

$$|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma| = \prod_{i=2}^{k}\|\mathbf{P}_{i-1}^\perp\mathbf{x}_i\|^2 = \prod_{i=2}^{k}[1-R^2_{i\cdot 1,\ldots,(i-1)}]. \tag{B.12}$$

For an arbitrary $i\in\{1,\ldots,k-1\}$, we consider the matrix $\tilde{\mathbf{X}}_\gamma=[\mathbf{x}_1\cdots\mathbf{x}_{i-1}\ \mathbf{x}_{i+1}\cdots\mathbf{x}_k\ \mathbf{x}_i]$, which is obtained by permuting the columns of $\mathbf{X}_\gamma$. For computing $|\tilde{\mathbf{X}}_\gamma^\top\tilde{\mathbf{X}}_\gamma|$, we apply the same technique like in (B.12). The fact that $|\tilde{\mathbf{X}}_\gamma^\top\tilde{\mathbf{X}}_\gamma|=|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|$ leads to

$$|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma| = [1-R^2_{\varsigma(k)\cdot\varsigma(1),\ldots,\varsigma(k-1)}]\prod_{j=2}^{k-1}[1-R^2_{\varsigma(j)\cdot\varsigma(1),\ldots,\varsigma(j-1)}]. \tag{B.13}$$

The identity in (42) is proven by combining (B.8), (B.11) and (B.13). We notice from (34) that $D_\gamma(\mathbf{y};\mathbf{Q}_3)=\|\mathbf{X}_\gamma^\top\mathbf{y}\|^2/|\mathbf{X}_\gamma^\top\mathbf{X}_\gamma|^{1/k}$, and by using (B.12), we get (43). □

## Appendix C. Proofs of the main results within Section 3.3

**Proof of Lemma 3.1.** For $i\in\{1,2,3\}$, we define $\mathbf{M}_i=\mathbf{X}_k(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{Q}_i(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{X}_k^\top/(|\mathbf{Q}_i|/|\mathbf{X}_k^\top\mathbf{X}_k|)^{1/k}$, and by applying a well-known result [30, p. 439], we have

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_i)] &= \mathbb{E}[\mathbf{y}^\top]\mathbf{M}_i\mathbb{E}[\mathbf{y}]+\tau\mathrm{Tr}[\mathbf{M}_i] \\ &= \boldsymbol{\beta}_{k\text{-}1}^\top\mathbf{X}_{k\text{-}1}^\top\mathbf{M}_i\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1}+\tau\frac{\mathrm{Tr}[(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}\mathbf{Q}_i]}{(|\mathbf{Q}_i|/|\mathbf{X}_k^\top\mathbf{X}_k|)^{1/k}},\end{aligned} \tag{C.1}$$

where $\mathrm{Tr}[\cdot]$ denotes the trace operator. When $\mathbf{Q}=\mathbf{Q}_1$, we compute (C.1) by making use of techniques similar with those employed to derive (B.1):

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)] &= \|\mathbf{P}_k\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1}\|^2+\tau\mathrm{Tr}[\mathbf{I}] \\ &= \|(\mathbf{P}_{k\text{-}1}+\omega^{-1}\mathbf{P}_{k\text{-}1}^\perp\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_{k\text{-}1}^\perp)\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1}\|^2+\tau k \\ &= \|\beta_{k-1}\|^2+\tau k.\end{aligned}$$

Hence, the identity in (45) is proven. Next we focus on some results that will be useful when evaluating (C.1) for $\mathbf{Q}=\mathbf{Q}_2$ and $\mathbf{Q}=\mathbf{Q}_3$. First notice from (B.5) that $|\mathbf{X}_k^\top\mathbf{X}_k|=\omega$. Moreover, we have from (B.2) that

$$\begin{aligned}\mathrm{Tr}[(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}] &= \mathrm{Tr}[(\mathbf{X}_{k\text{-}1}^\top\mathbf{P}_{\mathbf{x}_k}^\perp\mathbf{X}_{k\text{-}1})^{-1}]+(\mathbf{x}_k^\top\mathbf{P}_{k\text{-}1}^\perp\mathbf{x}_k)^{-1} \\ &= \mathrm{Tr}[(\mathbf{X}_{k\text{-}1}^\top\mathbf{P}_{\mathbf{x}_k}^\perp\mathbf{X}_{k\text{-}1})^{-1}]+\omega^{-1} = k-2+2\omega^{-1}.\end{aligned} \tag{C.2}$$

The identity above is deduced by taking into account that the eigenvalues of $(\mathbf{X}_{k\text{-}1}^\top\mathbf{P}_{\mathbf{x}_k}^\perp\mathbf{X}_{k\text{-}1})^{-1}$ are 1 and $\omega^{-1}$. The eigenvalue 1 has multiplicity $k-2$, while the eigenvalue $\omega^{-1}$ has multiplicity 1 [2, Eq. (25)]. These results together with (C.1) and some algebra yield (46) and (47):

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)] &= [\|\mathbf{X}_k^{\#}\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1}\|^2+\tau\mathrm{Tr}[(\mathbf{X}_k^\top\mathbf{X}_k)^{-1}]]|\mathbf{X}_k^\top\mathbf{X}_k|^{1/k}\\ &= [\|\boldsymbol{\beta}_{k\text{-}1}\|^2+\tau(k-2+2\omega^{-1})]\omega^{1/k},\end{aligned}$$

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)] &= [\|\mathbf{X}_k^\top\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1}\|^2+\tau\mathrm{Tr}[(\mathbf{X}_k^\top\mathbf{X}_k)]]|\mathbf{X}_k^\top\mathbf{X}_k|^{-1/k}\\ &= [\|\boldsymbol{\beta}_{k\text{-}1}\|^2+\tau k+(\mathbf{x}_k^\top\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1})^2]\omega^{-1/k}. \qquad \square\end{aligned}$$

**Proof of Proposition 3.4.**

(a) It follows from (45) and (46) that

$$\begin{aligned}&\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]\\ &\quad= \|\boldsymbol{\beta}_{k\text{-}1}\|^2[\omega^{1/k}-1]+\tau[k(\omega^{1/k}-1)-2\omega^{1/k-1}(\omega-1)]\\ &\quad= \tau(\omega^{1/k}-1)\left[\frac{\|\boldsymbol{\beta}_{k\text{-}1}\|^2}{\tau}+k-2\omega^{1/k-1}\,\frac{\omega-1}{\omega^{1/k}-1}\right]\\ &\quad= \tau(\omega^{1/k}-1)\left\{\frac{\|\boldsymbol{\beta}_{k\text{-}1}\|^2}{\tau}-\left[2\,\frac{1-\omega^{-1}}{1-\omega^{-1/k}}-k\right]\right\}.\end{aligned}$$

We can now infer the conclusion within point (a) of Proposition 3.4 by noticing that $\omega^{1/k}-1<0$ for $\alpha\in(0,\pi/2)$ and, additionally, $2(1-\omega^{-1})/(1-\omega^{-1/k})-k$ is a decreasing function of $\alpha$.

(b) The identities in (45) and (47) prove that

$$\begin{aligned}&\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]\\ &\quad= \mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)](\omega^{-1/k}-1)+(\mathbf{x}_k^\top\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1})^2\omega^{-1/k} \qquad \text{(C.3)}\end{aligned}$$

$$\leq (\|\boldsymbol{\beta}_{k\text{-}1}\|^2+\tau k)(\omega^{-1/k}-1)+\|\boldsymbol{\beta}_{k\text{-}1}\|^2(1-\omega)\omega^{-1/k} \qquad \text{(C.4)}$$

It is evident from (C.3) that $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$ cannot be negative because both $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)](\omega^{-1/k}-1)$ and $(\mathbf{x}_k^\top\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1})^2\omega^{-1/k}$ are non-negative. Note that (C.4) is obtained by applying the Cauchy–Schwarz inequality [30, p. 258]:

$$(\mathbf{x}_k^\top\mathbf{X}_{k\text{-}1}\boldsymbol{\beta}_{k\text{-}1})^2 \leq \|\mathbf{x}_k^\top\mathbf{X}_{k\text{-}1}\|^2\|\boldsymbol{\beta}_{k\text{-}1}\|^2=(1-\omega)\|\boldsymbol{\beta}_{k\text{-}1}\|^2.$$

The inequality in (48) is a straightforward consequence of (C.4). □

**Proof of Proposition 3.5.** First we give three auxiliary results which will be instrumental for the main proof.

- *Result* #1 [36, p. 348]. Let $\mathbf{N}\in\mathbb{R}^{k\times k}$ be a symmetric matrix whose eigenvalues are $\nu_1,\ldots,\nu_k$. Also, let $\boldsymbol{\theta}\in\mathbb{R}^k\backslash\{\mathbf{0}\}$. Then

  $$\nu_{\min}\|\boldsymbol{\theta}\|^2\leq\boldsymbol{\theta}^\top\mathbf{N}\boldsymbol{\theta}\leq\nu_{\max}\|\boldsymbol{\theta}\|^2, \qquad \text{(C.5)}$$

  where $\nu_{\min}=\min_{1\leq i\leq k}\nu_i$ and $\nu_{\max}=\max_{1\leq i\leq k}\nu_i$.
- *Result* #2. The arithmetic–geometric–harmonic mean inequalities [21, p. 27] applied to the eigenvalues of $\mathbf{X}_k^\top\mathbf{X}_k$:

  $$\lambda_{\min}\leq\mathcal{H}_\lambda\leq\mathcal{G}_\lambda\leq\mathcal{A}_\lambda\leq\lambda_{\max}, \qquad \text{(C.6)}$$

  where $\lambda_{\min}=\min_{1\leq i\leq k}\lambda_i$ and $\lambda_{\max}=\max_{1\leq i\leq k}\lambda_i$.
- *Result* #3. If $\mathbf{X}_{k\text{-}1}^\top\mathbf{X}_{k\text{-}1}=\mathbf{I}$, $\|\mathbf{x}_k\|=1$ and $\alpha\in(0,\pi/2]$ is the principal angle between $\langle\mathbf{X}_{k\text{-}1}\rangle$ and $\langle\mathbf{x}_k\rangle$, then the eigenvalues of $\mathbf{X}_k^\top\mathbf{X}_k$ satisfy the inequalities

  $$1-\cos\alpha\leq\lambda_{\min}\leq\lambda_{\max}\leq1+\cos\alpha. \qquad \text{(C.7)}$$

*Proof*: Let $\mathbf{b}=\mathbf{X}_{k\text{-}1}^\top\mathbf{x}_k$ and $\mathbf{B}=\left[\begin{smallmatrix}\mathbf{0}&\mathbf{b}\\ \mathbf{b}^\top&0\end{smallmatrix}\right]$. The equality $\mathbf{B}=\mathbf{X}_k^\top\mathbf{X}_k-\mathbf{I}$ is evident, and it implies that the eigenvalues of $\mathbf{B}$ are $\lambda_1-1,\ldots,\lambda_k-1$. For $i\in\{1,\ldots,k\}$, if $\mathbf{v}_i$ is the eigenvector of $\mathbf{X}_k^\top\mathbf{X}_k$ associated with $\lambda_i$, then $\mathbf{v}_i$ is also the eigenvector of $\mathbf{B}$ associated with $\lambda_i-1$. With the convention that $\mathbf{b}=[b_1,\ldots,b_{k-1}]^\top$ and $\mathbf{v}_i=[\nu_{1,i},\ldots,\nu_{k,i}]^\top$, we have

$$(\lambda_i-1)\mathbf{v}_i=\begin{bmatrix}\mathbf{0}&\mathbf{b}\\ \mathbf{b}^\top&0\end{bmatrix}\mathbf{v}_i=\begin{bmatrix}b_1\nu_{k,i}\\ \vdots\\ b_{k-1}\nu_{k,i}\\ \sum_{j=1}^{k-1}b_j\nu_{j,i}\end{bmatrix}.$$

The identities $\|\mathbf{v}_i\|^2=1$ and $\|\mathbf{b}\|^2=\cos^2\alpha$ together with the Cauchy–Schwarz inequality [30, p. 258] yield

$$\begin{aligned}(\lambda_i-1)^2&=\nu_{k,i}^2\|\mathbf{b}\|^2+\left(\sum_{j=1}^{k-1}b_j\nu_{j,i}\right)^2\\ &\leq\nu_{k,i}^2\|\mathbf{b}\|^2+\|\mathbf{b}\|^2\sum_{j=1}^{k-1}\nu_{j,i}^2=\cos^2\alpha,\end{aligned}$$

which implies $1-\cos\alpha\leq\lambda_i\leq1+\cos\alpha$ for all $i\in\{1,\ldots,k\}$. □

*Main inequalities*:

(a) From (49) and (50), we get

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]=\boldsymbol{\beta}_k^\top\mathbf{L}\boldsymbol{\beta}_k+\tau k(\mathcal{G}_\lambda/\mathcal{H}_\lambda-1),$$

where $\mathbf{L}=\mathcal{G}_\lambda\mathbf{I}-\mathbf{X}_k^\top\mathbf{X}_k$. Observe that the smallest eigenvalue of $\mathbf{L}$ is

$$\ell_{\min}=\mathcal{G}_\lambda-\lambda_{\max}, \qquad \text{(C.8)}$$

and the largest eigenvalue of $\mathbf{L}$ is

$$\ell_{\max}=\mathcal{G}_\lambda-\lambda_{\min}. \qquad \text{(C.9)}$$

By making use of (B.5), it is easy to check that

$$\mathcal{G}_\lambda=\sin^{2/k}\alpha. \qquad \text{(C.10)}$$

The steps of the proof for the inequalities in (52) are outlined below. At each step, we indicate which result is used in demonstration.

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]&\overset{\text{(C.6)}}{\geq}\boldsymbol{\beta}_k^\top\mathbf{L}\boldsymbol{\beta}_k\\ &\overset{\text{(C.5)}}{\geq}\|\boldsymbol{\beta}_k\|^2\ell_{\min}\\ &\overset{\text{(C.8)}}{=}\|\boldsymbol{\beta}_k\|^2(\mathcal{G}_\lambda-\lambda_{\max})\\ &\overset{\text{(C.10)}}{=}\|\boldsymbol{\beta}_k\|^2(\sin^{2/k}\alpha-\lambda_{\max})\\ &\overset{\text{(C.7)}}{\geq}\|\boldsymbol{\beta}_k\|^2(\sin^{2/k}\alpha-\cos\alpha-1),\end{aligned}$$

$$\begin{aligned}\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]&\overset{\text{(C.6)}}{\leq}\boldsymbol{\beta}_k^\top\mathbf{L}\boldsymbol{\beta}_k+\tau k(\mathcal{G}_\lambda/\lambda_{\min}-1)\\ &\overset{\text{(C.5)}}{\leq}\|\boldsymbol{\beta}_k\|^2\ell_{\max}+\tau k(\mathcal{G}_\lambda/\lambda_{\min}-1)\\ &\overset{\text{(C.9)}}{=}\|\boldsymbol{\beta}_k\|^2(\mathcal{G}_\lambda-\lambda_{\min})+\tau k(\mathcal{G}_\lambda/\lambda_{\min}-1)\\ &\overset{\text{(C.7)}}{\leq}\|\boldsymbol{\beta}_k\|^2(\mathcal{G}_\lambda+\cos\alpha-1)+\tau k[\mathcal{G}_\lambda/(1-\cos\alpha)-1]\end{aligned}$$

$$\overset{(C.10)}{=} \|\boldsymbol{\beta}_k\|^2(\sin^{2/k}\alpha+\cos\alpha-1)+\tau k\left(\frac{\sin^{2/k}\alpha}{1-\cos\alpha}-1\right).$$

(b) By subtracting (49) from (51), we obtain

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]=\boldsymbol{\beta}_k^\top\mathbf{M}\boldsymbol{\beta}_k+\tau k(\mathcal{A}_\lambda/\mathcal{G}_\lambda-1),$$

where $\mathbf{M}=(\mathbf{X}_k^\top\mathbf{X}_k)^2/\mathcal{G}_\lambda-\mathbf{X}_k^\top\mathbf{X}_k$. Let us consider the mapping $\phi(z)=z^2/\mathcal{G}_\lambda-z$, which is defined for all $z\in\mathbb{R}$. The eigenvalues of $\mathbf{M}$ are $\mu_i=\phi(\lambda_i)$, $i\in\{1,\ldots,k\}$. The inequalities in (C.7), together with the well-known properties of $\phi(\cdot)$, guarantee that $\mu_{\max}$, the maximum eigenvalue of $\mathbf{M}$, has the property: $\mu_{\max}\le\max\{\phi(1-\cos\alpha),\phi(1+\cos\alpha)\}$. Because $\phi(1+\cos\alpha)-\phi(1-\cos\alpha)=2\cos\alpha(2\sin^{-2/k}\alpha-1)>0$, the following inequality holds true:

$$\mu_{\max}\le\phi(1+\cos\alpha). \tag{C.11}$$

Since the parabola defined by $\phi(\cdot)$ attains its minimum when $z=\mathcal{G}_\lambda/2$, it is obvious that $\mu_{\min}$, the minimum eigenvalue of $\mathbf{M}$, cannot be smaller than $\phi(\mathcal{G}_\lambda/2)$. So,

$$\mu_{\min}\ge\phi(\mathcal{G}_\lambda/2). \tag{C.12}$$

The inequality above can be improved by observing for $\alpha\ge\pi/3$ that $1-\cos\alpha\ge\mathcal{G}_\lambda/2$ for all $k\ge2$. In this case,

$$\mu_{\min}\ge\phi(1-\cos\alpha). \tag{C.13}$$

Similarly with the chain of inequalities for $\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_2)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]$, we write

$$\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]\overset{(C.6)}{\ge}\boldsymbol{\beta}_k^\top\mathbf{M}\boldsymbol{\beta}_k\overset{(C.5)}{\ge}\|\boldsymbol{\beta}_k\|^2\mu_{\min}.$$

From the inequality above we get (53) and (54) by employing (C.12) and (C.13). Then we focus on the proof of (55):

$$\begin{aligned}
\mathbb{E}[D_{\gamma_k}(\mathbf{y};\mathbf{Q}_3)-D_{\gamma_k}(\mathbf{y};\mathbf{Q}_1)]&\overset{(C.6)}{\le}\boldsymbol{\beta}_k^\top\mathbf{M}\boldsymbol{\beta}_k+\tau k(\lambda_{\max}/\mathcal{G}_\lambda-1)\\
&\overset{(C.5)}{\le}\|\boldsymbol{\beta}_k\|^2\mu_{\max}+\tau k(\lambda_{\max}/\mathcal{G}_\lambda-1)\\
&\overset{(C.7)}{\le}\|\boldsymbol{\beta}_k\|^2\mu_{\max}+\tau k[(1+\cos\alpha)/\mathcal{G}_\lambda-1]\\
&\overset{(C.11)}{\le}\|\boldsymbol{\beta}_k\|^2\phi(1+\cos\alpha)+\tau k[(1+\cos\alpha)/\mathcal{G}_\lambda-1]\\
&\overset{(C.10)}{=}\|\boldsymbol{\beta}_k\|^2\left[\frac{(1+\cos\alpha)^2}{\sin^{2/k}\alpha}-(1+\cos\alpha)\right]+\tau k\left(\frac{1+\cos\alpha}{\sin^{2/k}\alpha}-1\right). \qquad\square
\end{aligned}$$

## References

[1] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control AC-19 (1974) 716–723.

[2] R. Behrens, L. Scharf, Signal processing applications of oblique projection operators, IEEE Transactions on Signal Processing 42 (1994) 1413–1424.

[3] A. Björck, G. Golub, Numerical methods for computing angles between linear subspaces, Mathematics of Computation 27 (1973) 579–594.

[4] L. Breiman, D. Freedman, How many variables should be entered in a regression equation?, Journal of the American Statistical Association 78 (1983) 131–136

[5] P. Djuric, Asymptotic MAP criteria for model selection, IEEE Transactions on Signal Processing 46 (1998) 2726–2735.

[6] C. Giurcăneanu, Estimation of sinusoidal regression model by stochastic complexity, in: P. Grünwald, P. Myllymäki, I. Tabus, M. Weinberger, B. Yu (Eds.), Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday, TICSP Series, vol. 38, Tampere International Center for Signal Processing, Tampere, Finland, 〈http://www.cs.tut.fi/~tabus/TICSP_38_17.4.08.pdf〉, 2008, pp. 229–249.

[7] C. Giurcăneanu, S. Razavi, New insights on stochastic complexity, in: Proceedings of the Eusipco 2009, the 17th European Signal Processing Conference, Glasgow, Scotland, UK, pp. 2475–2479.

[8] P. Grünwald, The Minimum Description Length Principle, MIT Press, 2007.

[9] M. Hansen, B. Yu, Model selection and the principle of minimum description length, Journal of the American Statistical Association 96 (2001) 746–774.

[10] M. Hansen, B. Yu, Minimum description length model selection criteria for generalized linear models, in: D. Goldstein (Ed.), Science and Statistics: A Festchrift for Terry Speed, Institute of Mathematical Statistics, Lecture Notes-Monograph Series, vol. 40, 2002, pp. 145–164.

[11] A. Hanson, P.C.W. Fu, Applications of MDL to selected families of models, in: P. Grünwald, I. Myung, M. Pitt (Eds.), Advances in Minimum Description Length: Theory and Applications, MIT Press, 2005, pp. 125–150.

[12] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference, and Prediction, second ed., Springer, 2009.

[13] C. Hurvich, C.L. Tsai, Regression and time series model selection in small samples, Biometrika 76 (1989) 297–307.

[14] S. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, 1993.

[15] S. Kay, Conditional model order estimation, IEEE Transactions on Signal Processing 49 (2001) 1910–1917.

[16] S. Kay, Exponentially embedded families—new approaches to model order estimation, IEEE Transactions on Aerospace and Electronic Systems 41 (2005) 333–345.

[17] E. Liski, Normalized ML and the MDL principle for variable selection in linear regression, in: E. Liski, J. Isotalo, J. Niemelä, S. Puntanen, G. Styan (Eds.), Festschrift for Tarmo Pukkila on his 60th Birthday, University of Tampere, 2006, pp. 159–172.

[18] E. Liski, A. Liski, Minimum description length model selection in Gaussian regression under data constraints, in: B. Schipp, W. Krämer (Eds.), Statistical Inference, Econometric Analysis and Matrix Algebra, Festschrift in Honour of Götz Trenkler, Springer, 2009, pp. 201–208 〈http://www.springerlink.com/content/n32t4671713w87g0/fulltext.pdf〉.

[19] K. Mardia, J. Kent, J. Bibby, Multivariate Analysis, Academic Press, 1979.

[20] T. McWhorter, L. Scharf, Cramer–Rao bounds for deterministic modal analysis, IEEE Transactions on Signal Processing 41 (1993) 1847–1866.

[21] D. Mitrinovic, P.M. Vasic, Analytic Inequalities, Springer Verlag, 1970.

[22] G. Qian, H. Künsch, Some notes on Rissanen's stochastic complexity, IEEE Transactions on Information Theory 44 (1998) 782–786.

[23] J. Rissanen, Modeling by shortest data description, Automatica 14 (1978) 465–471.

[24] J. Rissanen, Fisher information and stochastic complexity, IEEE Transactions on Information Theory 42 (1996) 40–47.

[25] J. Rissanen, MDL denoising, IEEE Transactions on Information Theory 46 (2000) 2537–2543.

[26] J. Rissanen, Information and Complexity in Statistical Modeling, Springer, 2007.

[27] T. Roos, P. Myllymäki, J. Rissanen, MDL denoising revisited, IEEE Transactions on Signal Processing 57 (2009) 3347–3360.

[28] D. Schmidt, E. Makalic, MML invariant linear regression, in: A. Nicholson, X. Li (Eds.), AI 2009: Advances in Artificial Intelligence, 22nd Australasian Joint Conference, Melbourne, Australia, December 1–4, 2009. Proceedings, Lecture Notes in Computer Science, vol. 5866, Springer2009, pp. 312–321.

[29] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (1978) 461–464.

[30] G. Seber, A Matrix Handbook for Statisticians, John Wiley & Sons, 2008.

[31] G. Seber, A. Lee, Linear Regression Analysis, Wiley-Interscience, 2003.

[32] A.K. Seghouane, Asymptotic bootstrap corrections of AIC for linear regression models, Signal Processing 90 (2010) 217–224.

[33] A.K. Seghouane, S. Amari, The AIC criterion and symmetrizing the Kullback–Leibler divergence, IEEE Transactions on Neural Networks 18 (2007) 97–106.

[34] A.K. Seghouane, M. Bekara, A small sample model selection criterion based on Kullback's symmetric divergence, IEEE Transactions on Signal Processing 52 (2004) 3314–3323.

[35] R. Shibata, Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, The Annals of Statistics 8 (1980) 147–164.

[36] P. Stoica, R. Moses, Spectral Analysis of Signals, Prentice-Hall, 2005.