Automatic uncertainty evaluation for determining the number of components in nested models

L. Martino*, R. San Millán-Castillo†, E. Morgado†

* Universitá degli studi di Catania, Italy. Email: luca.martino@unict.it

† Universidad Rey Juan Carlos, Fuenlabrada, Madrid, Spain.

Emails: roberto.sanmillan@urjc.es, eduardo.morgado@urjc.es

Abstract

In this work, we propose and examine two procedures for constructing intervals that capture the uncertainty associated with determining the effective number of components in model selection problems. The output of these methods is an interval (defined by two integer bounds) representing plausible values for the number of components. A detailed discussion is provided on the connection between the proposed approaches and the widely-used information criteria in the literature. Notably, the methods do not rely on the availability of a likelihood function, making them broadly applicable across various domains such as regression, classification, feature and/or order selection, clustering, and dimensionality reduction. These techniques leverage geometric properties of the error curve to construct the intervals. Extensive experiments on both synthetic and real-world datasets demonstrate the effectiveness and practical utility of the proposed procedures. Additionally, MATLAB code is provided to facilitate adoption by practitioners and researchers.

Keywords: Model Selection; uncertainty; elbow detection; information criteria; AIC; BIC.

1 Introduction

Model selection has become a fundamental task in contemporary signal processing, machine learning, and statistical analysis. See the following works [1], [2], [3] and [4] as examples showing the wide range of application through different fields. Moreover, in recent decades, uncertainty quantification and sensitivity analysis have emerged as highly relevant research topics across various scientific fields [5, 6, 7, 8]. Particularly important is the scenario of

nested models, i.e., a family of models of different complexity where the number of parameters can grow (i.e., the dimension of the vector of parameters can grow, building more complex models). This scenario appears frequently in different real-world applications: for instance, the order selection in polynomial regression or autoregressive schemes [9], [10], feature selection [11], clustering [12], change point detection [13], and dimension reduction just to name a few [14],[15]. Other relevant examples in signal processing are the estimation of the number of signal sources [16] and the so-called structured parameter selection [17]. Note that, throughout this work, the terms variables, components, features, and parameters are used interchangeably to refer to the elements of a model.

In the literature, three primary classes of methods are commonly employed to infer the complexity of nested models. The first class is formed by the cross-validation (CV) techniques [18, 19] or similar strategies [20, 9]. The second class is the so-called probabilistic statistical measures, formed by two main sub-families: the information criteria (IC)[21, 12, 22, 23] and the marginal likelihood approach (a.k.a., Bayesian evidence) used in Bayesian inference [24, 4, 25]. Related schemes can also be found [26, 27, 28]. The third class comprises methods grounded in geometric considerations, such as automatic 'elbow' or 'knee-point' detectors [29, 30, 31, 32].

In [29], The authors demonstrate that the automatic elbow detectors proposed in the literature can be reformulated as a specific instance of an information criterion. Various information criteria (IC) differ in their choice of the slope parameter λ, which governs the penalization of model complexity [33] (Table 1 provides different special cases of IC). Another recent approach, known as the Spectral Information Criterion (SIC), considers the entire spectrum of possible values for the penalization parameter λ, thereby encompassing other information criteria with linear complexity penalization as special cases. SIC also returns a confidence measure of the proposed solution [34, 35]. Similarly, other measures of reliability and confidence in the results in the context of elbow detection have been discussed [36]. These quantities attempt to show how 'safe' is the solution, in terms of possible information lost by constructing a 'too' parsimonious model.

In this work, the main contribution is twofold: we extend one of the derivations proposed in [29] and the SIC derivation, in order to provide an *interval* of indices corresponding to different possible models. Namely, the output of the proposed methods is an interval of possible number of components (defined by two specified integer values). This interval embeds the uncertainty associated with the decision in a black-box manner, contingent upon the specific analysis and the observed data. In the first proposed method, the underlying idea employs geometrical considerations, and it is inspired by the concept of maximum area-under-the-curve (AUC) in receiver operating characteristic (ROC) curves [37, 38] and by the derivation of the well-known Gini index [39, 40, 41, 42]. We also show the relationship of the proposed procedure with an alternating optimization considering conditioned information criteria. The second proposed method employs the confidence measure provided by SIC in order to obtain an interval of possible models as a final solution. The proposed schemes can also be applied in

more general contexts than the standard IC, even where a probabilistic model is not considered and likelihood function is not defined. Numerical experiments, with artificial and real data, show very promising results. Related Matlab code is also provided.[1]

2 Framework and background

2.1 The error curve $V(k)$

In numerous real-world applications, we desire to infer a vector of parameters $\boldsymbol{\theta}_k = [\theta_1, ..., \theta_k]^\top$ of dimension k given a data vector $\mathbf{y} = [y_1, ..., y_N]^\top$. An observation model that induces a likelihood function $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is usually available [25, 24]. The discrete variable k, which denotes the dimension of the parameter vector $\boldsymbol{\theta}_k$, is often unknown and must be inferred from the observed data $\mathbf{y}$. For example, k may represent the number of clusters in a clustering problem or the order of a polynomial in a nonlinear regression task. In these application problems, a non-increasing error curve can be computed,

$$V(k) : \mathbb{N} \to \mathbb{R}, \qquad k = 0, 1, 2, ..., K.$$

The function $V(k)$ can be any metric that characterizes the performance of the system. For simplicity and without loss of generality, we are considering an integer variable k with an increasing step of one unit. Clearly, it can be easily generalized. when a likelihood function is given, a usual choice of $V(k)$ is

$$V(k) = -2\log(\ell_{\mathtt{max}}), \quad \text{where} \quad \ell_{\mathtt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k),$$

e.g., as in [22, 43, 44, 25]. Alternatively, $V(k)$ can be directly defined as the mean square error MSE), or the mean absolute error (MAE) or transformations of them (e.g., as $\log$ MSE). However, any other fitting measure can be considered.

Remark. For the sake of simplicity and without loss of generality, we assume that $\min V(k) = V(K) = 0$. Clearly, this condition can always be obtained with a simple subtraction, $V'(k) = V(k) - \min V(k)$. See Figure 1 for an example of $V(k)$ (dashed line).

2.2 A general expression for several information criteria (IC)

In this section, we outline a general formulation for various information criteria (IC) and the underlying principles of this approach. Typically, the curve $V(k)$ is constructed as a non-increasing function. Graphical examples can be found in Figures 1 and 2(a). A well-established

[1] `http://www.lucamartino.altervista.org/PUBLIC_INTERVALS_CODE.zip`

method in the literature involves incorporating a linear penalty to account for model complexity,

$$C(k) = V(k) + \lambda k, \quad \lambda > 0, \tag{1}$$

where the slope of the complexity penalization term is denoted as λ. Since $V(k)$ is non-increasing and the penalty term λk increases with respect to k, the cost function $C(k)$ will exhibit at least one minimum. See Figure 2(b).
The objective is to obtain $k^* = \arg\min C(k)$ as the index of the 'optimal' model, serving as an estimator for the number of components in the nested model. Table 1 summarizes some relevant special cases of IC, considering the possible choice of error curve $V(k)$ and of the parameter λ. Each one has been derived in different contexts and with different assumptions. In the literature, there are also other IC with other analytical forms (e.g., with non-linear penalty terms), but they are not as widely employed as the IC with the analytical form in Eq. (1).

Table 1: Relevant examples of information criteria in the literature, with the corresponding choices of $V(k)$ and λ. Note that N denotes the number of data points and $\ell_{\max}$ is the maximum value reached by the likelihood function.

Information criterion	**Choice of λ**	$V(k)$
Bayesian-Schwarz (BIC) [43]	$\log N$	$-2\log \ell_{\max}$
Akaike (AIC) [44]	2	$-2\log \ell_{\max}$
Hannan-Quinn (HQIC) [45]	$\log(\log(N))$	$-2\log \ell_{\max}$
Universal Automatic Elbow Detector (UAED) [29]	$V(0)/K$	any
Spectral IC (SIC) [34]	all	any

3 Geometric-based design of the interval

In this section, we present a novel technique for obtaining an interval of indices in the context of an elbow detection problem, which encodes the uncertainty in model selection. More specifically, we extend one of the derivations of the universal automatic elbow detector, as presented in [29]. The underlying idea is inspired by the AUC approach in ROC curves for classification [37, 38], and the derivation of the Gini index in the economic field [39, 42].
The algorithm is based on the construction of three straight lines: the first one passing through the points $(0, V(0))$, to $(k_1, V(k_2))$, the second one passing through the points $(k_1, V(k_1))$ to $(k_2, V(k_2))$ and the last one passing through the points $(k_2, V(k_2))$ to $(K, 0)$, as shown in Figure 1 (clearly, $k_2 > k_1$). The goal is to minimize the area under this piece-linear approximation of the curve $V(k)$. The total area under this approximation is the sum of the two trapezoidal

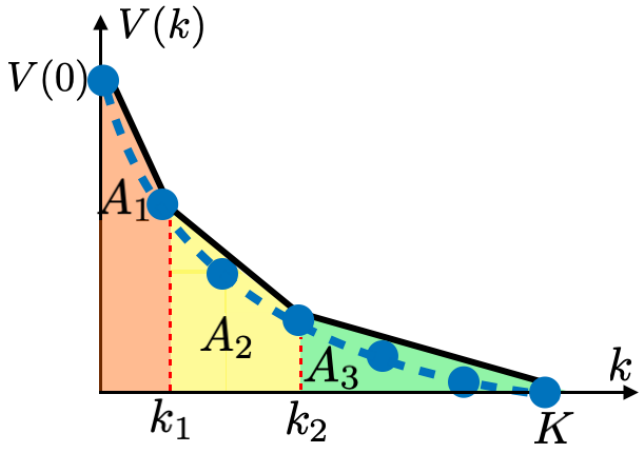


Figure 1: Example of error function $V(k)$ (dashed line) and the construction of the areas A_1, A_2, A_3, with three straight lines and $k_1 < k_2$.

areas (A_1 and A_2) and a triangular area (A_3), as depicted in Figure 1. Namely, we have

$$\begin{aligned} A_1 &= \frac{(V(0)+V(k_1))k_1}{2}, \\ A_2 &= \frac{(V(k_1)+V(k_2))(k_2-k_1)}{2}, \quad \text{and} \\ A_3 &= \frac{V(k_2)(K-k_2)}{2}, \end{aligned}$$

where we have used the assumption $V(K) = 0$. Hence the optimal interval which possibly includes the location of an "elbow" point is defined as

$$\begin{aligned} [k_1^*, k_2^*] &= \arg\min_{k_1,k_2} J(k_1,k_2), \\ &= \arg\min_{k_1,k_2} \{A_1 + A_2 + A_3\}, \\ &= \arg\min_{k_1,k_2} \Big\{k_1V(0) + k_2V(k_1) - k_1V(k_2) + KV(k_2)\Big\}, \\ &= \arg\min_{k_1,k_2} \Big\{k_1V(0) + k_2V(k_1) + (K-k_1)V(k_2)\Big\}, \quad k_1 < k_2. \end{aligned} \tag{2}$$

It is important to remark that solving this optimization above is straightforward because k_1 and k_2 belong to a discrete and finite set. Note that by construction, the elbow $k_{\texttt{UAED}}$ provided by UAED in [29] is always contained in the interval $[k_1, k_2]$, i.e., $k_{\texttt{UAED}} \in [k_1, k_2]$.

Relationship with information criteria. If we keep fixed k_2 as a constant, we can rewrite $J(k_1, k_2)$ as a function of only k_1 (given k_2), i.e.,

$$\begin{aligned} k_1^* &= \arg\min_{k_1} J(k_1|k_2) = \arg\min_{k_1} \Big\{ k_1 V(0) + k_2 V(k_1) + (K - k_1) V(k_2) \Big\}, \\ &= \arg\min_{k_1} \Big\{ k_2 V(k_1) + (V(0) - V(k_2)) k_1 + K V(k_2) \Big\}, \\ &= \arg\min_{k_1} \Big\{ V(k_1) + \frac{V(0) - V(k_2)}{k_2} k_1 \Big\}, \\ &= \arg\min_{k_1} \Big\{ C(k_1|k_2) \Big\}, \end{aligned} \tag{3}$$

where we have used that k_2 and K are considered constant. The last expression

$$\begin{aligned} C(k_1|k_2) &= V(k_1) + \frac{V(0) - V(k_2)}{k_2} k_1, \\ &= V(k_1) + \lambda(k_2) k_1, \end{aligned} \tag{4}$$

has the form of an information criterion where $V(k_1)$ plays the role of the error curve and the slope λ is here a function of k_2, i.e., $\lambda(k_2) = \frac{V(0) - V(k_2)}{k_2}$ (instead of being a constant). If $k_2 = K$ and since $V(K) = 0$, we recover the slope $\lambda = \frac{V(0)}{K}$ associated to the automatic elbow detector in [29], as shown in Table 1. Fixing k_1 as a constant, we can rewrite $J(k_1, k_2)$ as a function of only k_2 (given k_1), i.e.,

$$\begin{aligned} k_2^* &= \arg\min_{k_2} J(k_2|k_1) = \arg\min_{k_2} \Big\{ k_1 V(0) + k_2 V(k_1) + (K - k_1) V(k_2) \Big\} \\ &= \arg\min_{k_2} \Big\{ k_2 V(k_1) + (K - k_1) V(k_2) \Big\} \\ &= \arg\min_{k_2} \Big\{ V(k_2) + \frac{V(k_1)}{K - k_1} k_2 \Big\}, \\ &= \arg\min_{k_2} \Big\{ C(k_2|k_1) \Big\}, \end{aligned} \tag{5}$$

where we have used that k_1 and K are constant, and set

$$\begin{aligned} C(k_1|k_2) &= V(k_2) + \frac{V(k_1)}{K - k_1} k_2, \\ &= V(k_2) + \lambda(k_1) k_2, \end{aligned} \tag{6}$$

that, again, has the form of an information criterion where the slope λ is a function of k_1, i.e., $\lambda(k_1) = \frac{V(k_1)}{K - k_1}$, instead of a constant. If we set $k_1 = 0$, we recover again the slope $\lambda = \frac{V(0)}{K}$ associated to the automatic elbow detector in [29]. Then, thinking in an alternating

optimization approach, minimizing $J(k_1, k_2) = A_1 + A_2 + A_3$ can be interpreted as minimizing iteratively different conditioned information criteria, connected to each other by the different slopes of the complexity penalty.

4 SIC-based design of the interval

For simplicity, assume that $V(k)$ is strictly decreasing, so that $C(k)$ has a unique minimum. In [34], firstly the authors note that $\lambda \in [0, \lambda_{\texttt{max}}]$,

$$\lambda_{\texttt{max}} = \max_k \left[\frac{V(0) - V(k)}{k} \right], \quad \text{for } k = 1, ..., K. \tag{7}$$

Indeed, for $\lambda \geq \lambda_{\texttt{max}}$ we always obtain $k^* = \arg\min C(k) = 0$.[2] Generally, varying λ in $[0, \lambda_{\texttt{max}}]$, the position of the minimum changes k^*. Namely, the location of the minimum a function of λ,

$$k^*(\lambda) = \arg\min_k C(k, \lambda), \quad k^*(\lambda) : [0, \lambda_{\texttt{max}}] \to \{0, 1, 2..., K\}. \tag{8}$$

It is a non-increasing, piecewise constant function taking discrete values from 0 to K, where $k^*(0) = K$ and $k^*(\lambda) = 0$ for $\lambda \geq \lambda_{\texttt{max}}$. It is a piecewise constant function since, even changing λ, the value $k^*(\lambda)$ can remain unvaried. See Figure 2(a) for an example of function $V(k)$ and the corresponding cost function $C(k, \lambda)$ for a given value of λ. The resulting function $k^*(\lambda)$ is given in Figure 2(b).

As shown in Figure 2(b), different values of λ' and λ'' can yield the same minimum denoted as j, i.e., $k^*(\lambda') = j$ and $k^*(\lambda'') = j$, so that $k^*(\lambda)$ remains constant in certain pieces, each characterized by a specific length'. We can convert these 'lengths' into weights, associating to each index $j \in \{0, 1, .., K\}$ a normalized weight $\bar{w}_j$. A Monte Carlo procedure to compute these normalized weights $\bar{w}_j$ is given in Table 2. A more efficient alternative to the Monte Carlo approach is to use a fine grid for the values of λ, which is the strategy implemented in the provided code. Note that, by construction, we have $\bar{w}_0 = 0$.

In order to design an interval $[k_1^*, k_2^*]$ including the possible elbow of the curve (and encoding the related uncertainty), we define the cumulative sum of the first m weights, i.e.,

$$W_k = \sum_{j=1}^{k} \bar{w}_j,$$

with $0 < k \leq K$ and set

$$\begin{aligned} k_1^* &= \min\{k : \ W_k \geq \ell_1\}, \\ k_2^* &= \min\{k : \ W_k \geq \ell_2\}, \quad \ell_1 < \ell_2. \end{aligned} \tag{11}$$

[2] It is interesting to note the relationship between $\frac{V(0)-V(k)}{k}$ in Eq. (7) and $\lambda(k_2)$ in Eq. (4).

Table 2: Computation of the weights in the SIC method by Monte Carlo. Clearly, alternative quasi-Monte Carlo can be employed (using a fine grid of λ values).

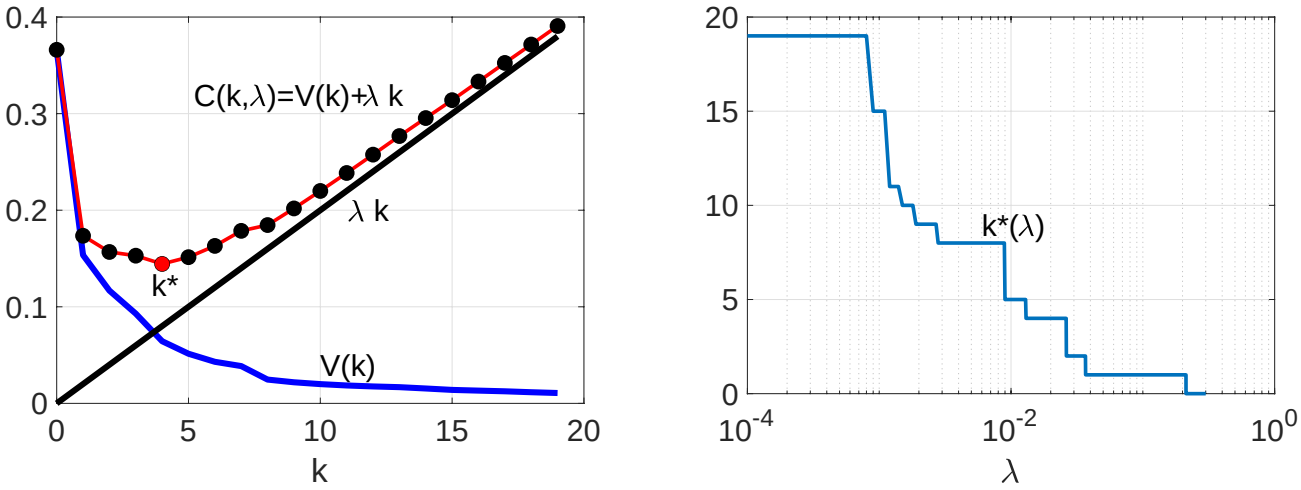
- Choose a value M (e.g., $M \geq 10^5$) and calculate $\lambda_{\texttt{max}} = \max_k \left[\frac{V(0)-V(k)}{k}\right]$.
- For $i = 1, ..., M$:
  1. Generate randomly $\lambda_i \sim \mathcal{U}([0, \lambda_{\texttt{max}}])$.
  2. Compute
$$k_i^* = \arg\min_k C(k, \lambda_i) = \arg\min_k \left[V(k) + \lambda_i k\right]. \tag{9}$$
- Return the frequency of the event $\{k_i^* = j\}$ for $j = 1, ..., K$, that is equivalent to return the weights
$$\bar{w}_j = \frac{\#\{k_i^* = j\}}{M}, \qquad j = 0, ..., K. \tag{10}$$

where ℓ_1, ℓ_2 play a similar role to a confidence level. After several empirical studies, we can assert that a very robust choice is given by $\ell_1 = 0.95$ and $\ell_2 = 0.995$. Safer intervals can be designed considering a more conservative value such as $\ell_2 = 0.999$. All the results in this work are obtained considering $\ell_2 = 0.995$.

5 Numerical experiments

In this section, we evaluate the two proposed constructions of the intervals across different settings, including experiments with artificial data (Sections 5.1 and 5.3), a synthetic curve $V(k)$ defined analytically (Section 5.2), and two real-world datasets (in the final two experiments). The MATLAB code used for the experiments is also made available.[3]

[3] `http://www.lucamartino.altervista.org/PUBLIC_INTERVALS_CODE.zip`



(a) (b)

Figure 2: **(a)** Example of function $V(k)$, the penalty λk (for a specific value of λ), and the resulting cost function $C(k, \lambda) = V(k) + \lambda k$ (shown with dots). **(b)** Example of piecewise constant function $k^*(\lambda)$ yielded by SIC in a log-λ scale.

5.1 Order selection in an auto-regressive model with a Laplacian noise

We generate a dataset of T pairs $\{t, y_t\}_{t=1}^T$, where t is an integer temporal index and the signal y_t is a scalar value for each t. We consider the following auto-regressive model,

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + ...\theta_k y_{t-k} + \epsilon_t, \quad \text{for } t = 1, ..., T, \tag{12}$$

where $\boldsymbol{\theta}_k = [\theta_1, \theta_2, ..., \theta_k]^\top$. We assume that ϵ_t is a Laplacian noise, i.e.,

$$p(\epsilon_t) = \frac{1}{2b} \exp\left(-\frac{|\epsilon_t - \mu|}{b}\right),$$

with zero mean, $\mu = 0$, and variance $\sigma_\epsilon^2 = 2b^2$. The goal is to infer the true order of the model $k_{\texttt{true}}$. This kind of problem is very frequent in signal processing, statistics, and machine learning [9, 10]. Note that it is possible to generate easily random samples from a Laplace density [46, Chapter 2]. Therefore, starting with null initial conditions, it is possible to generate data according to the model in Eq. (12). We test two levels of noise $\sigma_\epsilon = 0.5$, i.e., $b = 0.35$, with standard deviation $\sigma_\epsilon = 5$, i.e., $b = 3.53$. We also consider different numbers of data, $T \in \{200, 2000\}$.

In this example, we test 3 possible values of the order of the model, $k_{\texttt{true}} \in \{3, 5, 7\}$, where we have employed the following formula of the coefficients $\theta_i = (-1)^{i-1} \exp\{-0.3(i-1)\}$, to ensure that the system in Eq. (12) is stable. More specifically, we have

$$\begin{aligned}
k_{\texttt{true}} = 3 \implies & \theta_1 = 1,\ \theta_2 = -0.7408,\ \theta_3 = 0.5488; \\
k_{\texttt{true}} = 5 \implies & \theta_1 = 1,\ \theta_2 = -0.7408,\ \theta_3 = 0.5488,\ \theta_4 = -0.4066,\ \theta_5 = 0.3012; \\
k_{\texttt{true}} = 7 \implies & \theta_1 = 1,\ \theta_2 = -0.7408,\ \theta_3 = 0.5488,\ \theta_4 = -0.4066,\ \theta_5 = 0.3012, \\
& \theta_6 = -0.2231,\ \theta_7 = 0.1653.
\end{aligned}$$

Note that the last coefficients become smaller and smaller making more difficult their detection/estimation, especially with $\sigma_\epsilon = 5$. Hence, the scenario with $k_{\texttt{true}} = 7$ is more difficult in terms of estimation of the order of the model, especially with an high noise power. Given each combination of the values of σ_ϵ, T, and $k_{\texttt{true}}$, we generate the data $\mathbf{y} = [y_1, ..., y_T]$ according to the model (12). In all scenarios, we average the results with 10^3 independent runs, generating a new time series of T at each run. Moreover, in all the simulations, we consider $V(k) = -2\log(\ell_{\texttt{max}})$ with $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$ with $k \leq K$ (setting $K = 100$), where $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is induced by Eq. (12), in order to allow the comparison with other schemes in the literature, as shown in Tables 1 and 4. Note that, since we consider a Laplacian noise, $\ell_{\texttt{max}}$ is not provided with an analytic form. Indeed, it requires the knowledge of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_k = \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$, then $\ell_{\texttt{max}} = p(\mathbf{y}|\widehat{\boldsymbol{\theta}}_k)$. We consider the least squares estimation of the vector $\boldsymbol{\theta}_k$ as an approximation of $\widehat{\boldsymbol{\theta}}_k$. Figure 3 shows 50 examples of the curves $V(k)$ in different runs (for $k_{\texttt{true}} = 7$, $\sigma_\epsilon = 5$ and $T \in \{200, 2000\}$), jointly with the median curve (black solid line) and a yellow area between two black dashed lines, corresponding to the 98% of the empirical distribution provided by the 10^3 runs.

In Table 1, we provide the results of different information criteria: BIC [43], AIC [44], HQIC [34], and UAED [29]. The best results are remarked with a yellow colored cell. We can observe that BIC and UAED provide the best results in terms of correct-decision rate $p_A \in [0, 1]$, inferring the order of the model (i.e., the times that the method selects the correct order over the total number of simulations). Recall that the first proposed interval method is designed to extend the derivation of UAED. Table 4 shows the median intervals (obtained with the two methods introduced in the previous section) over the 10^3 runs in the different scenarios. Moreover, the table gives the rate $p_{\texttt{in}} \in [0, 1]$ of the intervals containing $k_{\texttt{true}}$ (clearly, including the extreme values of the interval). Namely, $p_{\texttt{in}}$ is the ratio of the number of times $k_{\texttt{true}}$ is inside the built interval over the total number of runs. This value $p_{\texttt{in}}$ plays the same role of the quantity $1 - \alpha$ where α is the confidence level in classical interval estimation. We can observe that both reach excellent performance. The best results in terms of shorter interval and $p_{\texttt{in}}$ are remarked with a green colored cell. Note that both methods seem to be virtually insensible to the noise power. The geometric-based approach always obtains good results even with a small number of data. Whereas the SIC-based approach seems to work better with a bigger number

of data, suffering more the case with smaller data (designing much longer intervals in these cases). We can see clearly this point looking to the median lengths $L_{\texttt{med}}$ of the intervals:

$$\textbf{Geometric-based} \Longrightarrow L_{\texttt{med}} \in \{9, 2, 11, 2, 4, 4, 4, 4, 6, 5, 6, 5\}$$
$$\textbf{SIC-based} \Longrightarrow L_{\texttt{med}} \in \{6, 0, 8, 0, 85, 0, 77, 0, 90, 4, 90, 4\}.$$

The SIC approach is able to return intervals with zero median lengths (maximum certainty) when $T = 2000$ (bigger number of data) and $k_{\texttt{true}} = 3$ and $k_{\texttt{true}} = 5$, but struggles when $T = 200$ where the median lengths are quite big such as 85, 77, 90 and 90, compared with the median lengths obtained with the geometric approach, respectively 4, 4, 6 and 6. In this sense, the geometric approach seems to be more robust.

Table 3: Summary of the correct-decision rate $p_A \in [0, 1]$ (averaged over 10^3 runs) for the IC methods, in the experiment of Section 5.1.

Scenario			**Method**			
$k_{\texttt{true}}$	σ_ϵ	T	**UAED**	**BIC**	**AIC**	**HQIC**
3	0.5	200	$p_A \approx 0.94$	$p_A \approx 0.97$	$p_A \approx 0.79$	$p_A \approx 0.73$
		2000	$p_A = 1$	$p_A = 1$	$p_A \approx 0.88$	$p_A \approx 0.89$
	5	200	$p_A \approx 0.96$	$p_A \approx 0.99$	$p_A \approx 0.78$	$p_A \approx 0.71$
		2000	$p_A = 1$	$p_A \approx 0.99$	$p_A \approx 0.89$	$p_A \approx 0.90$
5	0.5	200	$p_A \approx 0.89$	$p_A = 0.95$	$p_A \approx 0.78$	$p_A \approx 0.72$
		2000	$p_A = 1$	$p_A \approx 0.99$	$p_A \approx 0.84$	$p_A \approx 0.84$
	5	200	$p_A \approx 0.89$	$p_A \approx 0.97$	$p_A \approx 0.77$	$p_A \approx 0.68$
		2000	$p_A = 1$	$p_A \approx 0.99$	$p_A \approx 0.83$	$p_A \approx 0.83$
7	0.5	200	$p_A \approx 0.58$	$p_A \approx 0.30$	$p_A \approx 0.60$	$p_A \approx 0.56$
		2000	$p_A = 1$	$p_A \approx 0.98$	$p_A \approx 0.79$	$p_A \approx 0.81$
	5	200	$p_A \approx 0.58$	$p_A \approx 0.30$	$p_A \approx 0.58$	$p_A \approx 0.54$
		2000	$p_A = 1$	$p_A \approx 0.99$	$p_A \approx 0.78$	$p_A \approx 0.78$

Table 4: Performance of the designed intervals in Section 5.1. We provide **(a)** the median interval $I_{\texttt{med}}$ over all the runs, and **(b)** the rate $p_{\texttt{in}} \in [0,1]$ of the intervals containing $k_{\texttt{true}}$ (clearly, including the extremes of the interval). This value $p_{\texttt{in}}$ plays the same role of the quantity $1-\alpha$ where α is the confidence level (in classical interval estimation).

Scenario			**Method**	
$k_{\texttt{true}}$	σ_ϵ	T	**Geometric-based**	**SIC-based**
3	0.5	200	$I_{\texttt{med}} = [3,12]$ $p_{\texttt{in}} \approx 0.997$	$I_{\texttt{med}} = [3,9]$ $p_{\texttt{in}} = 1$
		2000	$I_{\texttt{med}} = [1,3]$ $p_{\texttt{in}} = 1$	$I_{\texttt{med}} = [3,3]$ $p_{\texttt{in}} = 1$
	5	200	$I_{\texttt{med}} = [3,14]$ $p_{\texttt{in}} \approx 0.995$	$I_{\texttt{med}} = [3,11]$ $p_{\texttt{in}} = 1$
		2000	$I_{\texttt{med}} = [1,3]$ $p_{\texttt{in}} = 1$	$I_{\texttt{med}} = [3,3]$ $p_{\texttt{in}} = 1$
5	0.5	200	$I_{\texttt{med}} = [2,6]$ $p_{\texttt{in}} \approx 0.995$	$I_{\texttt{med}} = [5,90]$ $p_{\texttt{in}} \approx 1$
		2000	$I_{\texttt{med}} = [1,5]$ $p_{\texttt{in}} = 1$	$I_{\texttt{med}} = [5,5]$ $p_{\texttt{in}} = 1$
	5	200	$I_{\texttt{med}} = [2,6]$ $p_{\texttt{in}} \approx 0.992$	$I_{\texttt{med}} = [5,82]$ $p_{\texttt{in}} = 1$
		2000	$I_{\texttt{med}} = [1,5]$ $p_{\texttt{in}} = 1$	$I_{\texttt{med}} = [5,5]$ $p_{\texttt{in}} = 1$
7	0.5	200	$I_{\texttt{med}} = [2,8]$ $p_{\texttt{in}} \approx 0.935$	$I_{\texttt{med}} = [3,93]$ $p_{\texttt{in}} \approx 0.983$
		2000	$I_{\texttt{med}} = [2,7]$ $p_{\texttt{in}} = 1$	$I_{\texttt{med}} = [3,7]$ $p_{\texttt{in}} = 1$
	5	200	$I_{\texttt{med}} = [2,8]$ $p_{\texttt{in}} \approx 0.927$	$I_{\texttt{med}} = [3,93]$ $p_{\texttt{in}} \approx 0.970$
		2000	$I_{\texttt{med}} = [2,7]$ $p_{\texttt{in}} = 1$	$I_{\texttt{med}} = [3,7]$ $p_{\texttt{in}} = 1$

5.2 Artificial curve $V(k)$ with known analytic form

In this section, we consider synthetic error curve $V(k)$ where we know its analytic form. In this way, we can observe some convergence properties of the different schemes and the ability to encode geometric invariant features of a curve $V(k)$. More specifically, the goal of this experiment is to analyze the behaviors of the elbow detectors [29, 30, 31, 32], of the index of effective number of variables (ENV) recently introduced in [36], and the two constructions of intervals proposed in this work, the geometric-based and the SIC-based approaches. We consider the function

$$V'(k) = e^{-0.1k}, \quad k = 0, 1, 2..., K,$$

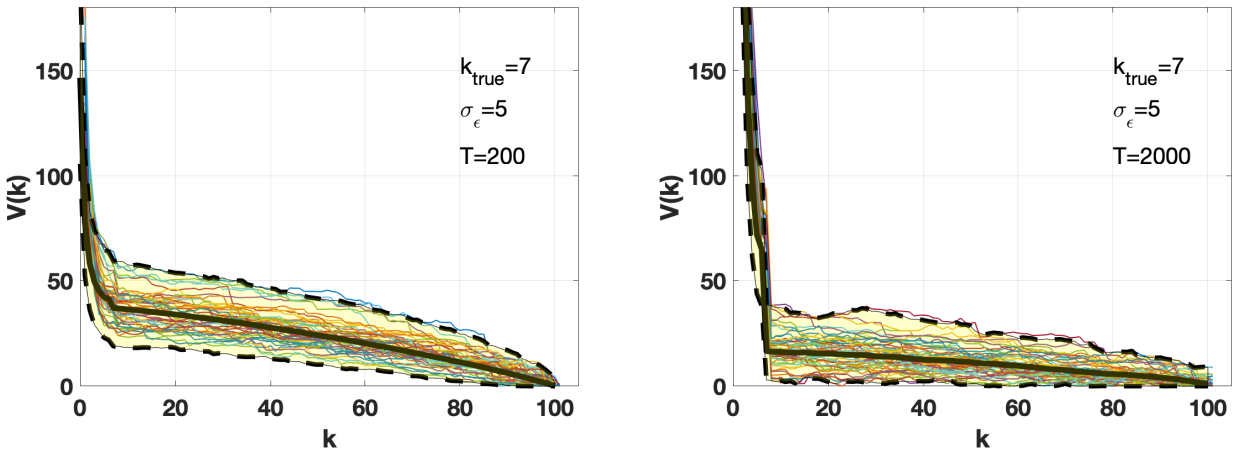


Figure 3: Examples of the curves $V(k)$ of Section 5.1 in 50 different runs for $k_{\texttt{true}} = 7$, $\sigma_\epsilon = 5$ and $T \in \{200, 2000\}$; the median curve is depicted with a black solid line, and the yellow area between two black dashed lines represents the 98% of the empirical distribution provided by the 10^3 runs.

and study different values of K. For each possible value of $K \in \{20, 50, 500, 5000, 10^4\}$, we define $V(k) = V'(k) - \min V'(k) = e^{-0.1k} - e^{-0.1K}$, so that $V(K) = 0$ in any case. We test the elbow detectors [29, 30, 31, 32], the ENV index and the constructions of the intervals, obtaining the results given in Table 5.

Table 5: Results in the synthetic experiment of Section 5.2.

Methods	$K = 20$	$K = 50$	$K = 500$	$K = 5000$	$K = 10^4$
Elbow detectors	8	16	39	62	69
ENV index	13.756	19.338	20.016	20.016	20.016
Geometric-based interval	[5,12]	[10,24]	[17,56]	[21,83]	[22,91]
SIC-based interval	[20,20]	[30,50]	[30,53]	[30,54]	[30,54]

We can observe that the detected elbows are always contained in the geometric-based intervals, as expected and stated in Section 3. However, the geometric-based intervals present an undesirable dependence on K, especially in the second extreme of the interval. On the other hand, the ENV index converges to the value $\bar{I}_{\texttt{ENV}} = 20.016$ as K grows, as expected [36]. The SIC-based intervals present the same stability property converging to the interval $[30, 54]$ as K grows.

5.3 Choice of the number of clusters

Let us consider a mixture of 5 bidimensional Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ where:

- $\boldsymbol{\mu}_1 = [3, 0]$, $\boldsymbol{\Sigma}_1 = [0.3, 0; 0, 2]$,
- $\boldsymbol{\mu}_2 = [14, 5]$, $\boldsymbol{\Sigma}_2 = [1.5, 0.7; 0.7, 1.5]$,
- $\boldsymbol{\mu}_3 = [-5, -10]$, $\boldsymbol{\Sigma}_3 = [1.5, 0.7; 0.7, 1.5]$,
- $\boldsymbol{\mu}_4 = [10, -10]$, $\boldsymbol{\Sigma}_4 = [1.5, 0; 0, 1.5]$;,
- and $\boldsymbol{\mu}_5 = [-5, 5]$, $\boldsymbol{\Sigma}_5 = [1, -0.8; -0.8, 1]$.

We generate 2500 simulated data from this mixture. Thus, the generated data form 5 disjoint clusters. In this experiment, we define $V(k) = \log\left[\sum_{j=1}^{k+1} \mathrm{var}(j)\right]$, where $\mathrm{var}(j)$ represents the inner variance of the j-th cluster. Each value of $\mathrm{var}(j)$ has been averaged over 200 runs, applying a k-means algorithm for defining the memberships to each cluster at each run. The case $k = 0$ corresponds to a unique, single cluster formed by all data. Hence, the total number of clusters is given by $k + 1$. We assume $K = 50$ as the maximum number of possible clusters. Note that with this choice of $V(k)$, we can apply only UAED whereas the rest of IC in Table 1 cannot be applied. In this experiment, UAED suggests the right number of clusters, 5. The geometric-based interval in this example is $[2, 6]$, whereas the SIC-based interval is $[5, 6]$. Both contain the right number of clusters, and the SIC-based interval is shorter.

5.4 Feature selection in a soundscape emotion real dataset

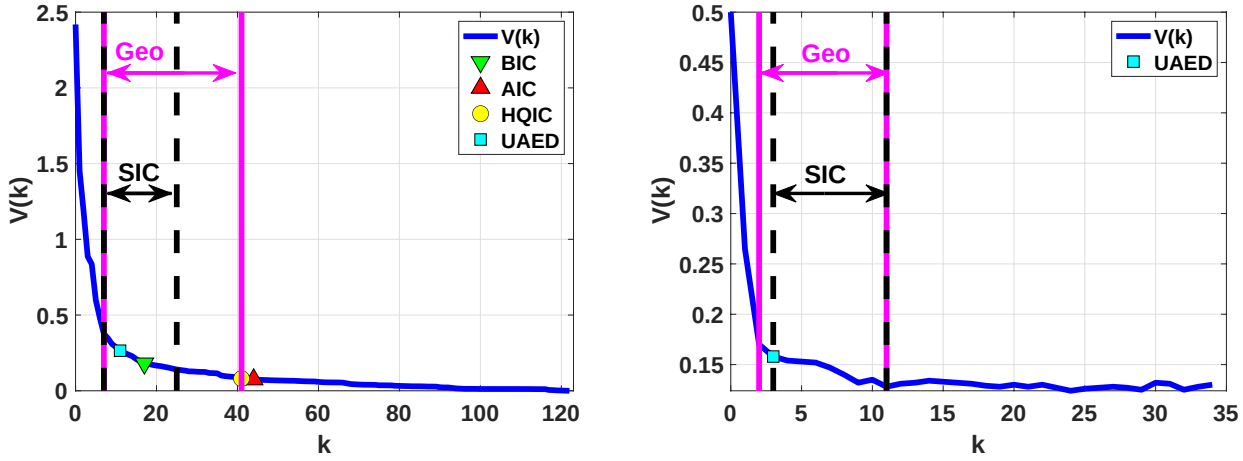
In this section, we address a variable selection problem in a regression setting using real-world data, specifically focusing on a soundscape emotion dataset. More specifically, a dataset of N pairs $\{\mathbf{x}_n, y_n\}_{n=1}^N$ is given, where each input vector $\mathbf{x}_n = [x_{n,1}, ..., x_{n,K}]$ is formed by K variables, and the outputs y_n's are scalar values. We assume $K \leq N$ and a linear measurement model,

$$y_n = \theta_0 + \theta_1 x_{n,1} + \theta_2 x_{n,2} + ...\theta_K x_{n,K} + \epsilon_n, \tag{13}$$

where ϵ_n is a Gaussian perturbation with zero mean and variance σ_ϵ^2, i.e., $\epsilon_n \sim \mathcal{N}(\epsilon|0, \sigma_\epsilon^2)$. In the soundscape emotion dataset analyzed for instance in [11], there are $K = 122$ features and $N = 1214$ number of data points. The output represents a variable defined as "arousal" in [11]. After ranking the 122 variables as suggested in [11], we set again $V(k) = -2\log(\ell_{\texttt{max}})$ where $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$ with $k \leq K$, and where the likelihood function $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is induced by Eq. (13). This choice of $V(k)$ allows the computation of AIC, BIC, and HQIC as well (see Table 1). The results of the different IC and the intervals built by the proposed schemes are given below:

- BIC suggests a model with 17 variables.
- AIC chooses 44 variables.
- HQIC selects a model with 41 variables.
- UAED suggests a model with 11 variables.
- The interval built with the geometric procedure is $[7, 41]$.
- The interval based on the SIC procedure is $[7, 25]$.

Note that that geometric-based interval contains all the IC with the exception of the result of AIC (44 variables). The SIC-based interval excludes also the solution provided by HQIC (41 variables). In this experiment, the SIC-based interval is shorter than the geometric-based interval. Both intervals are in line with other previous studies regarding this dataset and with the experts' recommendations in the literature, e.g., [11]. Figure 4(a) shows the corresponding curve $V(k)$ and summarizes with all the results provided by the IC and obtained intervals.



(a) (b)

Figure 4: $V(k)$ curves and results for the experiments in **(a)** Section 5.4 and **(b)** Section 5.5 with real data.

5.5 Feature selection in a classification problem with a nonalcoholic fatty liver disease real dataset

In this section, we consider an example of biomedical applications, which are nowadays extremely important in signal processing and machine learning [47, 48]. In [49], the authors study the most important variables for predicting patients at risk of developing nonalcoholic fatty liver disease. The dataset is formed by 1525 patients who attended the Cardiovascular Risk Unit of Mostoles University Hospital (Madrid, Spain) from 2005 to 2021. The authors in [49] employ a random forest (RF) algorithm as a classifier and rank the input variables, selecting the most relevant ones. The resulting 4 most important features are according to this raking: (a) insulin resistance, (b) ferritin, (c) serum levels of insulin, and (d) triglycerides. The authors in [49] employed cross-validation (CV) for finding the optimal number of features (that is 4) and this result was supported by the expert's opinions.
In this section, we have defined $V(k) = 1 - \text{accuracy}(k)$ as error curve that is given in Figure 4(b), using accuracy(k) obtained in [49] and after ranking the 35 variables as in [49]. Note that $V(0) = 0.5$ representing a completely random binary classification. Note that, even with this choice of the curve $V(k) = 1-\text{accuracy}(k)$, we can still apply UAED and SIC as shown in Table 1. The other information criteria cannot be applied in this context. However, we can obtain the intervals based on the geometric approach (which is related to the UAED derivation) and on the SIC approach. As shown in Figure 4(b), the resulting geometric interval is $[2, 11]$ and the SIC-based interval is $[3, 11]$. Both contain 4 variables which is exactly the result provided in [49], obtained by applying a cross-validation approach and supported by the experts' opinions.

6 Conclusions

In this work, we have proposed two alternative constructions for deriving intervals that capture the uncertainty associated with determining the number of components in nested models. The proposed approaches do not require knowledge of a likelihood function, making them broadly applicable across various domains, including regression and classification, feature and/or order selection, clustering, change point detection, and dimensionality reduction, among others. We have also extensively discussed the connection between our methods and widely used information criteria from the literature. Additionally, MATLAB code has been made available to facilitate the adoption of these methods by researchers and practitioners in applied settings. Extensive experiments on both synthetic and real data have demonstrated the strong performance of the proposed schemes. In particular, the results highlight that:

- Both procedures design intervals that contain the true number of components (or the number of components suggested by the experts) in more than 90% of the runs/realizations.

- The geometry-based procedure appears to be more robust to variations in the number of data points; however, its performance is influenced by the total number of components K.
- The SIC-based procedure tends to yield the most accurate results as the number of data points increases. Conversely, when the data size is limited, this approach often produces relatively wide intervals. A notable property of the SIC-based method is the convergence and invariance of the resulting interval as $K \to \infty$.
- Both approaches seems to exhibit a notable degree of robustness to the noise levels affecting the data.

The proposed schemes give particularly suitable solutions when the error curve $V(k)$ tends to be convex, as evidenced by the final two experiments involving real datasets and supported by our practical experience. It is also important to highlight that the proposed methods can be applied even in the absence of a likelihood function. Consequently, we argue that both procedures serve as (a) universal, (b) automatic, and (c) computationally efficient tools for quantifying the uncertainty inherent in model selection problems, without the need for resampling techniques or cross-validation schemes.

Acknowledgment

The work was partially supported by the project POLI-GRAPH, Grant PID2022-136887NB-I00 funded by MCIN/AEI/10.13039/501100011033.

References

[1] K. Aho, D. Derryberry, and T. Peterson, "Model selection for ecologists: the worldviews of AIC and BIC," *Ecology*, vol. 95, no. 3, pp. 631–636, 2014.

[2] A. Gupta and S. Das, "On efficient model selection for sparse hard and fuzzy center-based clustering algorithms," *Information Sciences*, vol. 590, pp. 29–44, 2022.

[3] N. L. Hjort and G. Claeskens, "Frequentist model average estimators," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 879–899, 2003.

[4] P. Stoica, X. Shang, and Y. Cheng, "The Monte-Carlo sampling approach to model selection: A primer [lecture notes]," *IEEE Signal Processing Magazine*, vol. 39, no. 5, pp. 85–92, 2022.

[5] M. Sensoy, L. M. Kaplan, S. Julier, M. Saleki, and F. Cerutti, "Risk-aware classification via uncertainty quantification," *Expert Systems with Applications*, vol. 265, p. 125906, 2025.

[6] G. Yang, "Multilayer neuroadaptive constraint-handling control architecture for a family of nonlinear systems with uncertainty compensation," *Information Sciences*, vol. 690, p. 121517, 2025.

[7] Q. Tan, Y. Wen, Y. Xu, K. Liu, S. He, and X. Bo, "Multi-view uncertainty deep forest: An innovative deep forest equipped with uncertainty estimation for drug-induced liver injury prediction," *Information Sciences*, vol. 667, p. 120342, 2024.

[8] J. Vicent Servera, L. Martino, J. Verrelst, J. P. Rivera-Caicedo, and G. Camps-Valls, "Multioutput feature selection for emulation and sensitivity analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

[9] W. L. Hung, E. S. Lee, and S. C. Chuang, "Balanced bootstrap resampling method for neural model selection," *Computers & Mathematics with Applications*, vol. 62, no. 12, pp. 4576–4581, 2011.

[10] Z. Zhu and S. Kay, "On Bayesian exponentially embedded family for model order selection," *IEEE Transactions on Signal Processing*, vol. 66, no. 4, pp. 933–943, 2017.

[11] R. San Millán, L. Martino, E. Morgado, and F. Llorente, "An exhaustive variable selection study for linear models of soundscape emotions: Rankings and Gibbs analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2460–2474, 2022.

[12] C. Cobos, H. Muñoz-Collazos, R.Urbano-Muñoz, M. Mendoza, E. León, and E. Herrera-Viedma, "Clustering of web search results based on the cuckoo search algorithm and balanced Bayesian information criterion," *Information Sciences*, vol. 281, pp. 248–264, 2014.

[13] M. Gupta, R. Wadhvani, and A. Rasool, "Comprehensive analysis of change-point dynamics detection in time series data: A review," *Expert Systems with Applications*, vol. 248, p. 123342, 2024.

[14] I. Gkioulekas and L. G. Papageorgiou, "Piecewise regression analysis through information criteria using mathematical programming," *Expert Systems with Applications*, vol. 121, pp. 362–372, 2019.

[15] P. Mukherjee, D. Parkinson, and A. R. Liddle, "A nested sampling algorithm for cosmological model selection," *The Astrophysical Journal Letters*, vol. 638, no. 2, p. L51, 2006.

[16] S. Beheshti and S. Sedghizadeh, "Number of source signal estimation by the mean squared eigenvalue error," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5694–5704, 2018.

[17] M. Jansen, "Information criteria for structured parameter selection in high-dimensional tree and graph models," *Digital Signal Processing*, vol. 148, p. 104437, 2024.

[18] E. Fong and C. Holmes, "On the marginal likelihood and cross-validation," *Biometrika*, vol. 107, no. 2, pp. 489–496, 2020.

[19] P. Stoica and Y. Selén, "Cross-validation rules for order estimation," *Digital Signal Processing*, vol. 14, pp. 355–371, 2004.

[20] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[21] T. Ando, "Predictive Bayesian model selection," *American Journal of Mathematical and Management Sciences*, vol. 31, no. 1-2, pp. 13–38, 2011.

[22] S. Konishi and G. Kitagawa, *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

[23] A. Van der Linde, "DIC in variable selection," *Statistica Neerlandica*, vol. 59, no. 1, pp. 45–56, 2005.

[24] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2004.

[25] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: an extensive review," *SIAM Review (SIREV)*, vol. 65, no. 1, pp. 3–58, 2023.

[26] D. P. Foster and E. I. George, "The risk inflation criterion for multiple regression," *The Annals of Statistics*, vol. 22, no. 4, pp. 1947–1975, 1994.

[27] C. L. Mallows, "Some comments on Cp," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

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

[29] E. Morgado, L. Martino, and R. S. Millan-Castillo, "Universal and automatic elbow detection for learning the effective number of components in model selection problems," *Digital Signal Processing*, vol. 140, p. 104103, 2023.

[30] A. J. Onumanyi, D. N. Molokomme, S. J. Isaac, and A. M. Abu-Mahfouz, "Autoelbow: An automatic elbow detection method for estimating the number of clusters in a dataset," *Applied Sciences*, vol. 12, no. 15, 2022.

[31] J. Zhang, P. Fu, F. Meng, X. Yang, J. Xu, and Y. Cui, "Estimation algorithm for chlorophyll-a concentrations in water from hyperspectral images based on feature derivation and ensemble learning," *Ecological Informatics*, vol. 71, p. 101783, 2022.

[32] D. Kaplan, "Knee point," 2024, MATLAB Central File Exchange. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/35094-knee-point

[33] A. Mariani, A. Giorgetti, and M. Chiani, "Model order selection based on information theoretic criteria: Design of the penalty," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2779–2789, 2015.

[34] L. Martino, R. S. Millan-Castillo, and E. Morgado, "Spectral information criterion for automatic elbow detection," *Expert Systems with Applications*, vol. 231, p. 120705, 2023.

[35] R. San Millán-Castillo, L. Martino, and E. Morgado, "A variable selection analysis for soundscape emotion modelling using decision tree regression and modern information criteria," *IEEE Access*, 2024.

[36] L. Martino, E. Morgado, and R. S. Millán-Castillo, "An index of effective number of variables for uncertainty and reliability analysis in model selection problems," *Signal Processing*, vol. 227, p. 109735, 2025.

[37] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, pp. 1–58, 2006.

[38] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[39] M. O. Lorenz, "Methods of measuring the concentration of wealth," *Publications of the American Statistical Association*, vol. 9, no. 70, pp. 209–219, 1905.

[40] L. Ceriani and P. Verme, "The origins of the Gini index: extracts from variabilitá e mutabilitá (1912) by Corrado Gini," *The Journal of Economic Inequality*, vol. 10, no. 3, pp. 421–443, 2012.

[41] S. Yitzhaki and E. Schechtman, *More Than a Dozen Alternative Ways of Spelling Gini*. Springer New York, 2013, pp. 11–31.

[42] S. Inoua, "Beware the Gini index! a new inequality measure," *preprint arXiv:2110.01741*, pp. 1–26, 2021.

[43] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[44] D. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. der Linde, "Bayesian measures of model complexity and fit," *J. R. Stat. Soc. B*, vol. 64, pp. 583–616, 2002.

[45] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 190–195, 1979.

[46] L. Martino, D. Luengo, and J. Míguez, "Independent random sampling methods," *Springer*, 2018.

[47] K. Ali, Z. A. Shaikh, and A. A. Khan, "Multiclass skin cancer classification using efficientnets-a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100034, 2022.

[48] V. Laghari, A. A.and Estrela and S. Yin, "How to collect and interpret medical pictures captured in highly challenging environments that range from nanoscale to hyperspectral imaging," *Curr Med Imaging*, pp. 1–20, 2022.

[49] R. Gárcia-Carretero, R. Holgado-Cuadrado, and O. Barquero-Pérez, "Assessment of classification models and relevant features on nonalcoholic steatohepatitis using random forest," *Entropy*, vol. 23, no. 6, 2021.