An index of effective number of variables for uncertainty and reliability analysis in model selection problems

Luca Martino[†], Eduardo Morgado[†], Roberto San Millán Castillo[†]

Abstract

An index of effective number of variables (ENV) is introduced for model selection in nested models. This is the case, for instance, when we have to decide the order of a polynomial function or the number of bases in a nonlinear regression, or choose the number of clusters in a clustering problem, or the number of feature in a variable selection application (to name few examples). It is inspired by the concept of maximum area under the curve (AUC) idea and the Gini index. The interpretation of the ENV index is identical to the effective sample size (ESS) indices with respect to a set of samples. The ENV index improves some drawback the elbow detectors described in the literature, and introduces different measures of uncertainty and reliability of the proposed solution. These novel reliability measures can be employed also jointly with the use different information criteria such as the well-known AIC and BIC. Comparisons with classical and recent schemes are provided in different experiments involving real datasets. Related Matlab code is given.

Keyword: Model selection, elbow detection, information criterion, Effective Sample Size (ESS), Gini index, uncertainty analysis, variable importance.

1 Introduction

Model selection is absolutely a fundamental task of scientific inquiry [1, 2, 3, 4]. It consists of selecting one model among many candidate models, given some observed data. We can distinguish three main frameworks in model selection. A first scenario is when completely different models are compared. The second setting is when several models defined by the same parametric family are considered (namely, these parameters of the model are tuned). The third setting is related to the previous one but, in this scenario, the family contains models of different complexity since the number of parameters can grow (i.e., the dimension of the vector of parameter grows, building more complex models). This last case, also known as *nested models*, is what we address in this work. Examples of model selection in nested models are the order selection in polynomial regression or autoregressive schemes, variable selection, clustering, dimension reduction, just to name a few [5, 6, 7, 8].

[†] Universidad Rey Juan Carlos, Campus de Fuenlabrada, Madrid

In the selection of nested models, the decision is driven by the so-called bias-variance trade-off: we have to choice a compromise between (a) the model performance and (b) the model complexity. Thus, the concept of selecting the best model is, in some sense, related to the mathematical definition of a 'good enough' model. Hence, the issue is to properly describe in terms of equations the colloquial expression 'good enough' [9].
In the literature, there are two main families of methods for model selection, that are composed by three main sub-families, as depicted in Figure 1. The first class is formed by *resampling methods*, where a main sub-family is given by the *bootstrap* and *cross-validation* (CV) techniques [10, 11, 12]. For simplicity, we focus here on CV. More specifically, CV is based on the splitting of the data in training and test sets. The training set is using to fit a model and the test set to evaluate it. However, the proportion of data to use in training (and/or in test) must be chosen by the user and can critically affect the results in terms of penalization of the model complexity. Moreover, the splitting process should be repeated several times, and the performance can be averaged over the runs (that is computationally costly).

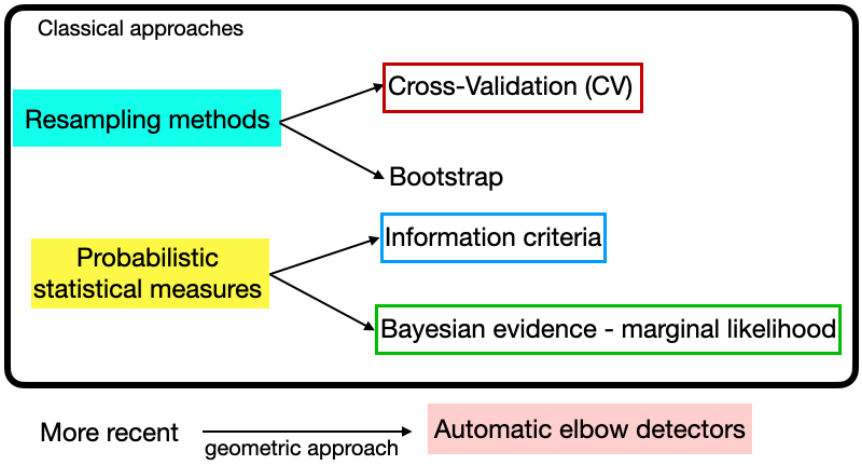


Figure 1: Classes of methods for model selection (standard ones and more recent approaches).

The second family is given by the so-called *probabilistic statistical measures*, which different employ rules for evaluating the different models considering directly the entire dataset (unlike in CV). This family is formed by two main sub-classes: the *information criteria* [13, 5, 14, 15] and the *marginal likelihood* approach (a.k.a., Bayesian evidence) used in Bayesian inference [16, 4, 17]. Some famous information criteria are the Akaike information criterion (AIC), which is based on entropy maximization principle [18], and the Bayesian information criterion (BIC) which is also a bridge with marginal likelihood approach, since BIC is derived as an approximation of the marginal likelihood [19]. Other very similar or completely equivalent schemes can be also found [20, 21, 22].
All the information criteria (IC) consider the sum of a performance measure and a penalty of the complexity. More specifically, they employ a linear penalization of the model complexity, where a parameter λ represents a positive slope value of this penalty. They differ for the *slope* λ of this linear penalization term (see, for instance, Table 1). The choice of this slope is justified by different theoretical derivations, each one with several assumptions and approximations. Clearly,

the results depend on this choice [23]. Some considerations regarding consistency of IC can be found in the survey [24]. The last approach is based on the marginal likelihood, that is employed in Bayesian inference for model selection purposes. The model penalization in the marginal likelihood is induced by the choice of the prior densities [25, 4].

Therefore, in the three main approaches described above, (a) CV, (b) information criteria and (c) marginal likelihood computation, we have always a degree of freedom (proportion of split data, slope λ, and choice of the prior densities) that affects the final results of the model selection. For this reason, other recent approaches have been also investigated in the literature, based on geometric considerations: some of them propose and automatic detection of an "elbow" or "knee-point" in an error curve [26, 27, 28, 29]. In [26], the authors provide four different and equivalent geometric derivations showing: (a) the elbow detectors in [26, 27, 28, 29] are equivalent (providing the same result), and (b) this result can be obtained as an optimization of an information criterion with an specific choice of λ (given in [26]), i.e., this geometric approach can be also expressed as an information criterion. On the other hand, another recent information criterion, called spectral information criterion (SIC) [23], has been also proposed in the literature: this criterion uses all the possible values of λ (thus also contains the rest of IC as special cases) and returns also a *reliability measure* of the proposed solution.

In this work, we extend one of the derivation proposed in [26] designing an index of effective number of variables (ENV). Note that, through all the work, we use the words *variables*, *components*, *features*, and/or *parameters* of a model as synonymous [30, 31]. The resulting index is inspired by the the concept of maximum area-under-the-curve (AUC) in receiver operating characteristic (ROC) curves [9, 32] and the Gini index [33, 34, 35, 36]. Moreover, the underlying idea and interpretation is exactly like the effective sample size (ESS) indices with respect to a set of samples [37, 38]. The ENV index improves the elbow detectors [26, 27, 28, 29] removing the dependence on the maximum number of variables K considered in the analysis.
Moreover, we introduce measures of reliability and uncertainty of the proposed results, related to ENV and similarly to what SIC delivers as well. They provide quantities associate to how 'safe' is the solution, in term of possible information lost (by constructing a 'too' parsimonious model). It is important to remark that the novel reliability measures can be employed also for different information criteria, such as AIC and BIC (to name a few), not only when the elbow detectors is applied. In order to define these reliability measures, we also introduce some variable importance measures (see also the sensibility analysis in [39]) that are in some way related to other famous concepts already proposed in literature, such as the Shapley values and feature importance ideas [40, 41].
The rest of work is structured as follows. In Section 2, we define the main notation and recall some background material. The ENV index is derived and analyzed in Section 3. Additional properties of the ENV index are given in Section 4. Furthermore, in the same section, we introduce some reliability and uncertainty measures based on the derivation of ENV index. In Section 5, we show different numerical experiments. We provide some final conclusions in Section 6.

2 Framework and main notation

2.1 The error curve $V(k)$ as figure of merit

In many applications in signal processing and machine learning, we desire to infer a vector of parameters $\boldsymbol{\theta}_k = [\theta_1, ..., \theta_k]^\top$ of dimension k given a data vector $\mathbf{y} = [y_1, ..., y_N]^\top$. A likelihood function $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is usually available, and often derived from a related physical model [17, 16]. In many scenarios, also the dimension k is unknown and must be estimated as well. This is the case in numerous applications, for instance when k can represent (a) the number of clusters in a clustering problem, (b) the order of a polynomial in a non-linear regression problem, (c) the number of selected variables in a feature selection problem, (d) the main number of dimensions in a dimension reduction problem, to name a few. All these examples can be encompassed as special cases of a more general class of problems named *model selection with nested models*, with an increasing order of complexity (where the complexity is defined by the number of *components* k included in the model). In all these real-world application problems, a non-increasing *error function* (i.e., a performance metric that characterizes the performance of the system, such as a fitting measure),

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

can be obtained. Note that $k = 0$ represents the "no model" scenario (or the simplest possible model). For instance, in a polynomial regression, the value $V(0)$ could correspond to the variance of the data, i.e., the mean square error (MSE) error in prediction using a constant value θ_0 (equal to the mean of the data) as model.
Generally, more complex models (with more parameters, then $\boldsymbol{\theta}_k = [\theta_0, \theta_1, ..., \theta_k]^\top$ has an higher dimension k) provide smaller errors in prediction/classification. We consider the most complex model to have K more parameters/components (than the simplest case with only θ_0), i.e., $\boldsymbol{\theta}_K = [\theta_0, \theta_1, ..., \theta_K]^\top$. Without lost of generality, we are considering an integer variable k with increasing step of one unit (more general assumptions could be done). In a variable/feature selection problem, we assume that the order of the variable inside the vector $\boldsymbol{\theta}_k$ is well-chosen, i.e., $V(k)$ is built after ranking the variables/features (in a decreasing order of relevance) [8].

Examples of possible choices of $V(k)$ are the following:

- In the literature, when a likelihood function is given, a usual choice is

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

  e.g., as in [14, 19, 18, 17].

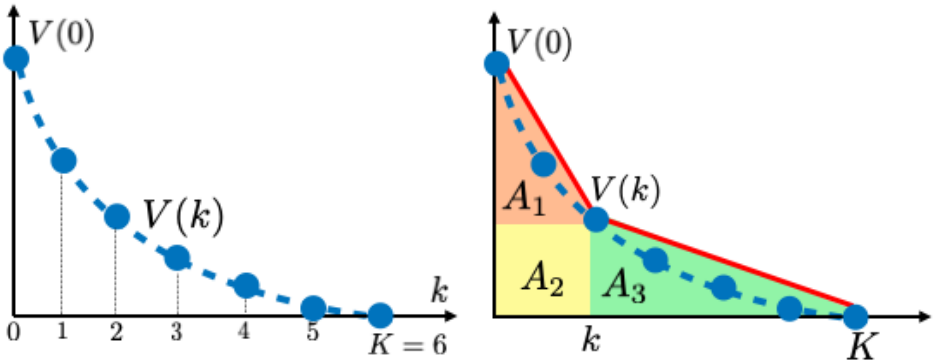
- $V(k)$ could be directly the MSE, and/or the mean absolute error (MAE) or transformations of them (as $\log$ MSE etc.). For instance, in the linear and additive Gaussian noise case, it is possible to show that the choice $V(k) = -2\log(\ell_{\texttt{max}})$ is equivalent to $V(k) = N \log \text{MSE}$ where the MSE represents also an estimation of the noise power in the system (e.g., see [42]).

- $V(k)$ can be any other error measure in classification or clustering, for instance. See Section 5.3 for an example.

Generally, $V(k)$ should be a *non-increasing* error curve, i.e., for any pair of non-negative integers n_1, n_2 such that $n_2 > n_1$, then we have $V(n_2) \leq V(n_1)$.[1] Indeed, $V(k)$ is a fitting term that decreases as the complexity of the model (given by the number k of parameters) grows. Therefore, we have $V(0) \geq V(k), \forall k$. Note that $V(k)$ plays the same role of the *Lorenz curve* in the definition of the Gini index [33, 34]. See Figure 2 for a graphical example.

Observe that $V(0)$ represents the value of the error function corresponding to the *simplest model*, for instance, a constant model in a regression problem, or a single cluster (for all the data) in a clustering problem. In some applications, the score function $V(k)$ should be also convex (as in Figure 2), i.e., the differences $V(k+1) - V(k)$ will decrease as k increases. This is the case of a variable selection problem, if the variables have been ranked correctly. However, this work does not require conditions regarding the concavity of $V(k)$.

Finally, just for the sake of simplicity and without loss of generality, we assume that $\min V(k) = V(K) = 0$. Clearly, This condition can be always obtained with a simple subtraction. For instance, given a generic non-increasing function $V'(k)$, we can define $V(k) = V'(k) - \min V'(k) = V'(k) - V'(K)$ so that $\min V(k) = 0$.



(a) (b)

Figure 2: **(a)** Example of error function $V(k)$ where $K = 6$, **(b)** Construction with two straight lines and the areas A_1, A_2 and A_3.

2.2 The universal automatic elbow detector

In this section, we briefly recall one of the derivations (all based on geometric arguments) of the universal automatic elbow detector (UAED) given in [26]. Here, we need to recall the derivation more connected to the AUC approach [9, 32]. The underlying idea is to extract geometric information from the curve $V(k)$ looking for a geometric "elbow" k_e in order to determine the optimal number of components (denoted as $k_e \in \{0, 1..., K\}$) to include in our model.
We consider the construction of two straight lines passing through the points $(0, V(0))$, to $(k, V(k))$

[1] This condition could be also relaxed. We keep it, for the sake of simplicity.

and $(k, V(k))$, to $(K, 0)$ as shown in Figure 2(b) (where $k \in \{0, 1, ..., K\}$). Hence, we have piecewise linear approximation of the curve $V(k)$ with these two straight lines. The goal is to minimize the area under this approximation. More specifically, the total area under the approximation is the sum of the two triangular areas (A_1 and A_3) and the rectangular area (A_2) in Figure 2(b). Namely, we have

$$A_1 = \frac{k(V(0) - V(k))}{2},\ A_2 = kV(k),\ A_3 = \frac{(K-k)V(k)}{2},$$

hence the optimal number of components k_e (location of the "elbow") is defined as

$$k_e = \arg\min_k \{A_1 + A_2 + A_3\}. \tag{1}$$

After some algebra, we arrive to the expression

$$k_e = \arg\min_k \left\{ V(k) + \frac{V(0)}{K} k \right\}, \quad \text{for } k = 1, ..., K, \tag{2}$$

where clearly we are assuming $V(0) \neq 0$ and $K \neq 0$.

Remark. It is important to remark that, since k belongs to a discrete and finite set, solving the optimization above is straightforward.

Remark. If $V(k)$ is convex, k_e is unique. Convexity of $V(k)$ is a sufficient but not necessary condition for the uniqueness of the solution.

Remark. If $V(k)$ is not convex, we can have several global minima. For instance, having M different minima, $k_1^*, k_2^*, ..., k_M^*$, the user can choose the best solution (within the M possible one) according to some specific requirement depending on the specific application. Here, we suggest to pick the most conservative choice, i.e., $k_e = \max k_j^*$, for $j = 1, ..., M$.

Relation with the information criteria. The cost function employed by UAED is

$$C(k) = V(k) + \frac{V(0)}{K} k, \tag{3}$$

$$= V(k) + \lambda k, \tag{4}$$

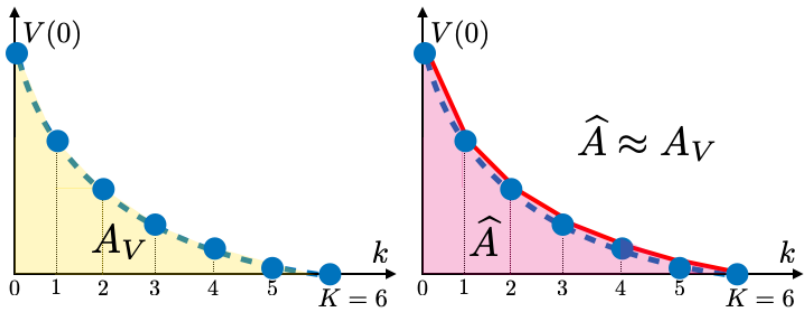
where the slope of the complexity penalization term is $\lambda = \frac{V(0)}{K}$. Therefore, this cost function has exactly the same form of the cost function used in the information criteria like BIC and AIC, i.e., with a linear penalization of the model complexity and selecting $\lambda = \frac{V(0)}{K}$. In AIC and or BIC, we have $V(k) = -2 \log \ell_{\texttt{max}}$. Therefore, when $V(k) = -2 \log \ell_{\texttt{max}}$, UAED can be interpreted as an information criterion with the particular choice of $\lambda = \frac{V(0)}{K}$. Table 1 summarizes this information.

3 An index of the effective number of variables (ENV)

UAED and the other elbow detectors provide very good performance in several different scenarios, as shown in [26]. This prove the strength of the geometric approach compared to other information

Table 1: Information criteria in the literature, with the corresponding choices of $V(k)$ and λ. Note that N denoted the number of data points and $\ell_{\max}$ is the maximum value reached by a likelihood function. ENV represents the scheme presented in this work.

Information criterion (IC)	**Choice of λ**	$V(k)$
Bayesian-Schwarz [19]	$\log N$	$-2\log \ell_{\max}$
Akaike [18]	2	$-2\log \ell_{\max}$
Hannan-Quinn [43]	$\log(\log(N))$	$-2\log \ell_{\max}$
Universal Automatic Elbow Detector [26]	$\frac{V(0)}{K}$	any
Spectral IC (SIC) [23]	all	any
Eff. num. of variables (ENV)	—	any



(a) (b)

Figure 3: **(a)** We can consider that $V(k)$ is like a sampled curve obtained from sampling - in a signal processing sense - a continuous function $V(x)$ where $x \in \mathbb{R}$ (shown in dashed line) is an auxiliary continuous variable. The continuous function $V(x)$ possibly do not exist, can be just a theoretical tool to define the area A_V. **(b)** In any case, we have accessed to $V(k)$, $k \in \mathbb{N}$, that allows to obtain the approximation $\widehat{A} \approx A_V$.

criteria proposed in the literature.

However, all the elbow detectors presented in the literature [26, 27, 28, 29] a dependence on the maximum of variables analyzed, i.e., K. This is due to the slope of the linear complexity penalty is $\lambda = \frac{V(0)}{K}$. See also Section 5.1, for a numerical example. In some applications, the number K is maximum value strictly defined by a limitation of the system/model. In other frameworks, K could be increased: if a new possible component is added even with a small impact on the decrease of error function $V(k)$, since the slope $\lambda = \frac{V(0)}{K}$ becomes smaller, the elbow detectors could suggest a bigger value as optimal number of components. On the other hand, if the model selection analysis is performed reducing the possible total number of components in advance, the elbow detectors would suggest a smaller value as optimal number of variables.[2] Moreover, in several applications, it is important also to obtain an uncertainty measure for the model selection, jointly with the elbow detection. In the next sections, we try to address these issues.

[2]Clearly, there is also a dependence of $V(0)$ and, more generally, the result depends on the choice of error curve $V(k)$, as well.

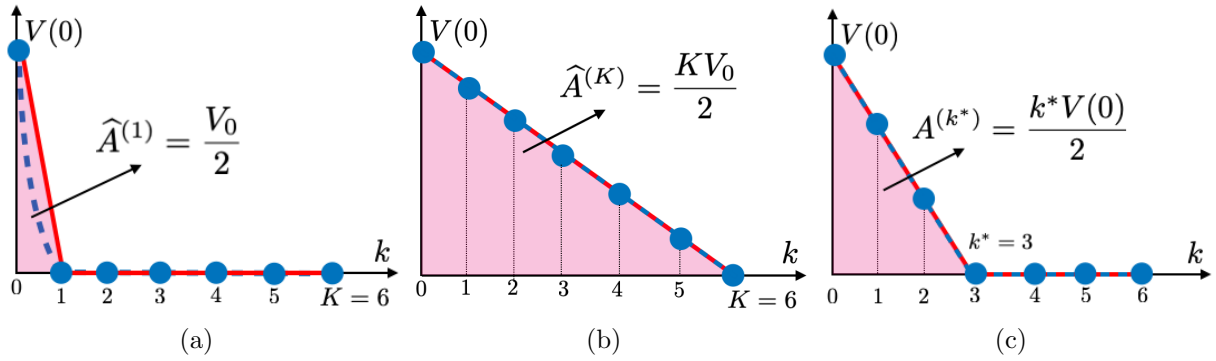


Figure 4: Special ideal cases. **(a)** $V(k)$ reaches zero already at $k = 1$. Only the first component is relevant, hence the optimal choice $k^* = k_e = 1$. **(b)** $V(k)$ is straight line connecting the point $(0, V(0))$ and $(K, 0)$. All variables contribute in the same way to the decay $V(k)$, hence the optimal choice $k^* = k_e = K$ (in figure $K = 6$). **(c)** The function $V(k)$ is a straight line passing through the points $(0, V(0))$ and $(k^*, 0)$, that is $V(k) = V(0) - \frac{V(0)}{k^*}k$, so that $V(k^*) = 0$ at some point $k^* < K$. Clearly, the point $k^* = k_e = 3$ is an optimal choice: the first 3 variables have the same contribution to the decay $V(k)$ and completely explain the drop.

3.1 Derivation of the ENV index

In this section, we extend the geometric approach employed UAED designing a new index that **(a)** reduces the dependence on K and **(b)** helps us to provide uncertainty measures with respect to the chosen elbow. The main idea is formed by two parts: **(Part-1)** to obtain a better approximation $\widehat{A}$ (w.r.t. the derivation of UAED) of the area under the curve $V(k)$, denoted as A_V and shown Figure 3(a), **(Part-2)** then we have to normalize the obtained value $\widehat{A}$ (considering some ideal scenario). We provide more details below.

(Part-1) A better approximation of the area under the function $V(k)$ can be easily obtained as sum of K trapezoidal pieces:

$$\begin{aligned}
\widehat{A} &= \sum_{k=0}^{K-1} \frac{V(k) + V(k+1)}{2}, \\
&= \frac{V(0)}{2} + \frac{V(1)}{2} + \frac{V(1)}{2} + \frac{V(2)}{2} + \ldots + \frac{V(K-1)}{2} + \frac{V(K)}{2}, \\
&= \frac{V(0) + V(K)}{2} + \sum_{k=1}^{K-1} V(k), \\
&= \frac{V(0)}{2} + \sum_{k=1}^{K-1} V(k), \qquad (5)
\end{aligned}$$

where, in the last equality, we have used the assumption that $V(K) = 0$ and the step of increase on axis k is 1 (without loss of generality). An example is depicted in Figure 3(b).

(Part-2) In order to design a *normalized* index effective of the number of components, we need

to describe an ideal scenario: let us assume that all entire drop of $V(k)$ is already reached in the first step, i.e., we have already $V(1) = 0$ and clearly $V(1) = V(2) = ...V(K) = 0$, as shown in Figure 4(a). In this case, the first variable/component is relevant whereas the rest of variables do not give any contribution to the decay of $V(k)$. Therefore, the correct decision as effective of the number of components is $k^* = k_e = 1$.[3] The approximated area under the curve $V(k)$ in this case is $\widehat{A}^{(1)} = \frac{V(0)}{2}$. Thus, we define the index of effective number of variables (ENV) as:

$$I_{\texttt{ENV}} = \begin{cases} 0, & \text{when } V(0) = 0, \\ \dfrac{\widehat{A}}{\widehat{A}^{(1)}} = \dfrac{2}{V(0)}\widehat{A}, & \text{when } V(0) \neq 0. \end{cases} \tag{6}$$

Note that, when $V(0) \neq 0$, we can then write

$$I_{\texttt{ENV}} = 1 + 2\sum_{k=1}^{K-1} \frac{V(k)}{V(0)}, \qquad (\text{for } V(0) \neq 0). \tag{7}$$

3.2 Behavior of $I_{\texttt{ENV}}$ in ideal cases

We check the behavior of the ENV index in different extreme and ideal cases:

- In the ideal scenario above shown in Figure 4(a), i.e., when $V(0) \neq 0$ but we have already $V(1) = 0$ and $V(1) = V(2) = ...V(K) = 0$, the correct decision is $k^* = 1$. Hence, note also that $\widehat{A} = \widehat{A}^{(1)}$. In this case, from (6) or from (7) we get $I_{\texttt{ENV}} = 1$, as desired.

- Another ideal scenario is when $V(k)$ is a linear straight line connecting the points $(0, V(0))$ and $(K, V(K))$, as shown in Figure 4(b). In this situation, all the variables contribute in the same way to the decay of $V(k)$ (i.e., each variable has the same influence on the error decrease), so that the correct decision is $k^* = K$ (all the components are relevant). The area under the curve $V(k)$ in this case is $\widehat{A}^{(K)} = \frac{KV(0)}{2}$. In this case, the equality $\widehat{A}^{(K)} = A_V$ also holds. With Eq. (6), we also obtain $\widehat{A} = \frac{KV(0)}{2}$ therefore the ENV index is $I_{\texttt{ENV}} = \frac{2}{V(0)}\frac{KV(0)}{2} = K$, exactly as expected.

- Another extreme case is when all the components are independent from the output y. In this situation, theoretically $V(k)$ should be a constant, $V(k) = V(0)$ for all k, namely $V(0) = V(1) = ...V(K) = 0$ (due to our assumption $V(K) = 0$, without losing any generality). Hence, the correct decision is $k^* = 0$ and the area under the function $V(k)$ is $\widehat{A}^{(0)} = 0$. in the second case, $\widehat{A} = 0$ so that $I_{\texttt{ENV}} = 0$, again as desired.

[3] If we suppose a continuous function that sampled at each step 1 generated the curve $V(k)$, the result $k^* = k_e = 1$ is optimal for us, even if the function decays faster than a linear decrease, *but only because* we have *not* access to sampled points between 0 and 1 (i.e., we have no information about the faster decay between 0 and 1); see Figure 4(a). Having more information, i.e., more points between 0 and 1 (hence an increase in k smaller than 1), then the optimal result would be $k^* < 1$ (if we have a decay as the dashed line in Figure 4(a)).

- More generally, consider at some k^* we have $V(k^*) = 0$ and the $V(k)$ is a straight line passing through the points $(0, V(0))$ and $(k^*, 0)$, that is $V(k) = V(0) - \frac{V(0)}{k^*}k$. Then,

$$\begin{aligned} I_{\texttt{ENV}} &= 1 + 2\sum_{k=1}^{k^*} \frac{V(0) - \frac{V(0)}{k^*}k}{V(0)}, \\ &= 1 + 2\sum_{k=1}^{k^*} \left(1 - \frac{k}{k^*}\right), \\ &= 1 + 2k^* - \frac{k^*(k^*+1)}{k^*}, \\ &= 1 + 2k^* - (k^*+1), \\ &= k^*. \end{aligned}$$

  Exactly as expected and desired. This can be also easily obtained looking Figure 4(c): indeed, in this scenario we can write $\widehat{A} = \frac{k^*V(0)}{2}$ and the ENV index will be $I_{\texttt{ENV}} = \frac{2}{V(0)}\frac{k^*V(0)}{2} = k^*$.

Hence, we have an index such that

$$0 \leq I_{\texttt{ENV}} \leq K. \tag{8}$$

Some additional properties and observations are given below.

4 Interpreting and using the ENV index

In this section, we introduce a property of the ENV index (given below) and which effects are shown in Section 5.1. An interesting behavior of $I_{\texttt{ENV}}$ in ideal scenarios is described below and depicted in Figure 5. More generally, we explain how interpreting and using $I_{\texttt{ENV}}$.

4.1 First considerations

Property 1. Let us consider a generic non-increasing function $V'(k)$, with $k = 0, ..., K$ and assume to have an horizontal asymptote, i.e., $\lim_{K\to\infty} V'(k) = C$. We can always define $V(k) = V'(k) - \min V'(k)$ so that $V(K) = 0$ for each possible K. Note that the ENV index in Eq. (7) depends on K (hence we use the notation here $I_{\texttt{ENV}} = I_{\texttt{ENV}}(K)$) but, when $K \to \infty$, we have that *stability property*, i.e.,

$$\lim_{K\to\infty} I_{\texttt{ENV}}(K) = \bar{I}_{\texttt{ENV}}, \tag{9}$$

i.e., the ENV index converges to a stable value $\bar{I}_{\texttt{ENV}}$, that represents a geometrical feature of the curve $V(k)$. This due to the fact that adding infinitesimal portions of area to A_V virtually does not change the values of A_V itself and $\widehat{A}$; see Figs. 3(a) and 3(b). Generally, this is not the

behavior of the elbow detectors [26, 27, 28, 29] in (non-ideal) real scenarios. See the numerical example in Section 5.1.

Property 2. Moreover, another property is that

$$I_{\texttt{ENV}} \geq 1, \quad \text{for } V(0) \neq 0, \tag{10}$$

as we can observe clearly from Eq. (7). This is due to the fact that k is a discrete variable with increasing step of 1, i.e., $k = 0, 1, ..., K$, and $V(0) = 0$ we have $V(k) = 0$ for all k by assumption.[4] Hence, if $V(k) = 0$ we have $I_{\texttt{ENV}} = 0$ as also shown in Eq. (6). If $V(0) \neq 0$, since k is discrete, the first zero can occur at least at $k = 1$ (and $V(1) = V(2) = ...V(K) = 0$ by assumption), and in that case $I_{\texttt{ENV}} = 1$. If the first zero happens at $k = 2$, $V(2) = 0$, depending on the type of decay that we have (i.e., considering the two differences $V(0) - V(1)$ and $V(1) - V(2)$) we will have $1 < I_{\texttt{ENV}} \leq 2$. The equality $I_{\texttt{ENV}} = 2$ is given is the decay is *linear* (see the concept of *variable importance* in Section 4.2). Thus, due to the discrete nature of k, we cannot have value $0 < I_{\texttt{ENV}} < 1$, but we can have $I_{\texttt{ENV}} = 0$, $I_{\texttt{ENV}} = 1$ and any value in the intervals $j - 1 \leq I_{\texttt{ENV}} \leq j$ for all $j = 2, ..., K$. In summary, the values that the ENV index are the following $I_{\texttt{ENV}} \in \{0, [1, K]\}$. This makes sense since there is any kind of decay, in $V(k)$, this means at least one component deserves to be considered in the model.

Choosing the number of components. Like an elbow detector, the ENV index can be employed for choosing the number of components in a model selection problem just defining

$$k^* = \lfloor I_{\texttt{ENV}} \rceil, \tag{11}$$

where $\lfloor a \rceil$ represents the rounding operation of replacing an arbitrary real number a by its nearest integer.

4.2 Reliability and uncertainty measures for the decision

Before to introduce some reliability and uncertainty measures, we describe some relevant ideal cases. In Figure 5, we consider four ideal scenarios where $V(k)$ is composed by two straight lines creating an elbow at $k_e = 14$. These ideal cases are shown with a blue solid line. A well-designed elbow detector as in [26, 27, 28, 29], always picks $k_e = 14$ since this is the position of the elbow (hence this is the correct result for an elbow detector). However, we can see that the four ideal scenarios the confidence in the choice $k_e = 14$ is different. In the lowest curve $V(k)$, the choice $k = 14$ (used as number of effective number of components) is clearly an optimal point, since $V(k)$ already reaches zero at $k = 14$. Whereas, the "quality" of this choice decreases as the blue line becomes closer and closer to the red line (that represents the ideal scenario where all components are equally relevant). This is also shown by the value $I_{\texttt{ENV}}$ (depicted red triangles) that becomes more and more distant to the elbow position k_e. These examples in Fig. 5 clarify the need of confidence and/or uncertainty measure related to the decision.

[4] We have access only to the "sampled" curve $V(k)$. We do not know $V(x)$ where $x \in \mathbb{R}$ that possibly defines a theoretical area A_V, as shown in Figure 3. If, in a specific problem, we have accessed to $V(x)$ than $I_{\texttt{ENV}}$ could take intermediate values also between 0 and 1.

A first attempt to provide a reliability measure has been given in the recent-proposed spectral information criterion (SIC) [23] which returns **(a)** a suggestion regarding the position of the elbow k_e (as the other information criteria) **(b)** but also provides a degree of *reliability* in the decision (as a confidence measure). SIC associates to its decision a number that is a sum of some normalized weights (e.g., often 0.95 and 0.99): closer to 1 (100%), the decision is more *safe*, meaning that we are confidence in discarding that $K - k_e$ variables/components in the construction of the model. The derivation used for UAED and extended here for obtaining the ENV index, can be also used for this purpose. The goal is to achieve a mathematical development that involves some "weights" as in the SIC scheme (i.e., moving closer to a SIC-like approach).

Variable importance. The underlying idea behind UAED and ENV is to associate a measure of *importance* of each variable/component, defined as a value proportional to its contribution to the decay of $V(k)$, i.e.,

$$w_k = V(k-1) - V(k), \quad k = 1, ..., K.$$

Hence, w_k represents the importance of k-th component. We can normalize these weights arriving to the following result:

$$\bar{w}_k = \frac{w_k}{\sum_{i=1}^{K-1} w_i} = \frac{V(k-1) - V(k)}{V(0)},$$

where we have used that

$$\sum_{i=1}^{K-1} w_i = V(0) - V(K) = V(0),$$

since $V(K) = 0$ by assumption. In the case of equally importance variables, given when $V(k)$ is a unique straight line as in Fig. 4(b) or the red line in Fig. 5, we have

$$V(k) = V(0) - \frac{V(0)}{K}k,$$
$$w_k = V(k-1) - V(k) = -\frac{V(0)}{K}(k-1) + \frac{V(0)}{K}k = \frac{V(0)}{K},$$

that is same value for all k (as expected) and, as consequence, $\bar{w}_k = \frac{1}{K}$ for all k. This definition of variable importance have been already applied with success in [39].

Cumulative importance (CI). We can compute the importance accumulated using the first (ranked) k components,

$$\mathrm{CI}(k) = \sum_{i=1}^{k} \bar{w}_i = \frac{\sum_{i=1}^{k} w_i}{V(0)} = \frac{V(0) - V(k)}{V(0)} = 1 - \frac{V(k)}{V(0)}$$

for $k = 0, ..., K$. Note that $0 \leq \mathrm{CI}(k) \leq 1$, $\mathrm{CI}(0) = 0$ and $\mathrm{CI}(K) = 1$. If the values zero is reached for a $k^* \leq K$, i.e., $V(k^*) = 0$, we have $\mathrm{CI}(k^*) = 1$. This cumulative sum of normalized weights can play the same roles of the cumulative sum of normalized weights in SIC (i.e., as a reliability/safety measure), although the computation of the weights follows a completely different procedure [23]. In the case of equally importance variables we have $\mathrm{CI}(k) = \frac{k}{K}$. The CI can be

considered an accuracy measure: closer to 1 we have more accuracy, i.e., we have safer and more reliable decisions. Hence, as a consequence, $1 - \mathrm{CI}(k)$ is an uncertainty measure (the amount of *lost importance*, lost choosing a model with only the first k components), and we can define also the cumulative uncertainty as

$$\mathrm{CU}(k) = 1 - \mathrm{CI}(k) = \frac{V(k)}{V(0)}. \tag{12}$$

We can rewrite the ENV index as sum of the cumulative uncertainties,

$$I_{\texttt{ENV}} = 1 + 2\sum_{k=1}^{K-1} \mathrm{CU}(k), \quad \text{for } V(0) \neq 0. \tag{13}$$

In Figure 5, from the bottom to the upper curve, the strengths of the 4 elbows at $k_e = 14$ are $\mathrm{CI}(k_e) = 1$, 0.87, 0.67, and 0.50, respectively.

Reliability of the decision. Other indicators of the reliability of the decision can be designed. Indeed, the value $I_{\texttt{ENV}}$ provides the *effective number* of variables/components in our model selection problem. Clearly, the decision of using $k < I_{\texttt{ENV}}$ variables is more *risky* (in terms of loss in performance) instead of using a model with $k > I_{\texttt{ENV}}$. In this sense, a suitable indicator for the reliability of the decision could be defined as

$$R_D = \min\left[1, \frac{k_e}{I_{\texttt{ENV}}}\right]. \tag{14}$$

Namely, any elbow position k_e such that $k_e - I_{\texttt{ENV}} > 0$ can be considered in a safe region (since we are using *more variables than the effective number of variables*), i.e., according to the ENV index, more parsimonious models could be chosen.

5 Numerical experiments

In this section, we test the ENV index in different frameworks, including experiments with artificial data (the first one) and two real datasets (the last two experiments). We will see that $I_{\texttt{ENV}}$ provides very good performance and presents more robustness with respect to other alternatives in the literature and we show the behavior of the reliability measures proposed above.[5]

5.1 Synthetic experiment where $V(k)$ is an analytic function

The goal of this example is to show the behaviors of the elbow detectors [26, 27, 28, 29] and the ENV index in a synthetic framework but that is not an ideal scenario (as in Figure 5 where an elbow is well-defined). We consider the function

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

[5]The Matlab code and datasets of the experiments are available at `http://www.lucamartino.altervista.org/PUBLIC_ENV_CODE.zip`.

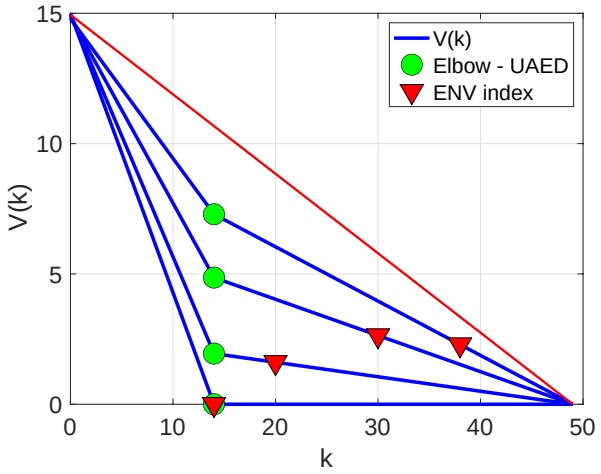


Figure 5: Ideal cases of fours curves $V(k)$ (blue solid lines) where an "elbow" is well-defined at $k_e = 14$. This is exactly the result that we obtain with the application of an elbow detector [26, 27, 28, 29] (green circles), whereas the results provided by the ENV index are shown by red triangles. Note that, in all cases, $I_{\texttt{ENV}} > k_e$.

We consider different value of K. For each possible value of $K \in \{20, 50, 500, 5000\}$, we define $V(k) = V'(k) - \min V'(k) = e^{-0.1k} - e^{-0.1K}$, so that $V(K) = 0$. We apply the elbow detectors [26, 27, 28, 29] and the ENV index obtaining the following results:

Methods	$K = 20$	$K = 50$	$K = 500$	$K = 5000$
Elbow detectors	8	16	39	62
ENV index	13.756	19.338	20.016	20.016
CI	0.58	0.78	0.98	0.99
R_D	0.58	0.83	1	1

hence, the ENV index converges to the value $\bar{I}_{\texttt{ENV}} = 20.016$. This stable value of the ENV index is an intrinsic geometric property of the analyzed curve $V(k)$. Moreover, we can observe that both the reliability indices are quite smaller than 1 when the elbow is smaller than $\bar{I}_{\texttt{ENV}} = 20.016$, and both approach 1 as the elbow position becomes greater and greater, as expected. We can also that the index of reliability of the decision R_D is more optimistic than the cumulative importance (CI).

5.2 Variable selection in a regression problem with real data

In a regression problem, we observe a dataset of N pairs $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$, where each input vector $\mathbf{x}_n = [x_{n,1}, \ldots, x_{n,K}]$ is formed by K variables, and the outputs y_n 's are scalar values [31]. We consider the case that $K \leq N$ and assume a linear observation model,

$$y_n = \theta_0 + \theta_1 x_{n,1} + \theta_2 x_{n,2} + \ldots \theta_K x_{n,K} + \epsilon_n, \tag{15}$$

where ϵ_n is a Gaussian noise with zero mean and variance σ_ϵ^2, i.e., $\epsilon_n \sim \mathcal{N}(\epsilon|0, \sigma_\epsilon^2)$. More specifically, in this real dataset [8, 44], there are $K = 122$ features and $N = 1214$ number of data

points $\mathbf{x}_i$. The dataset presents two outputs: "arousal" and "valence". In this section, we focus on the first output in the dataset ("arousal"). In this experiment, we can set $V(k) = -2\log(\ell_{\max})$ with $\ell_{\max} = \max_\theta p(\mathbf{y}|\boldsymbol{\theta}_k)$ with $k \leq K$, after ranking the 122 variables (see [8]), where the likelihood function $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is induced by the Eq. (15). Thus, here we can compare with other information criteria given in the literature, shown in Table 1 and a standard method based on the computation of p-values [45]- [46]. The ENV index returns a value of 12.74. Thus, the results provides by each method are given in the Table below:

Scheme	p-value	AIC	BIC	HQIC	UAED	SIC-95	SIC-99	**ENV**
k_e	71	44	17	41	11	7	17	**13**
CI	0.98	0.97	0.92	0.96	0.88	0.84	0.92	0.90
R_D	1	1	1	1	0.86	0.55	1	1
Ref.	[46]	[18]	[19]	[43]	[26]	[23]	[23]	here

The first line represents the methods, the second line gives the number of suggested variables, and the last line contains the corresponding references. In [8, Section 4-C], the results of that exhaustive analysis suggest that there are 7 very relevant variables (level 1 of [8, Section 4-C]), other 7 relevant variables (level 2 of [8, Section 4-C]) and other 2 variables in a level 3 of importance [8, Section 4-C], hence, totally 16 variables among very relevant, relevant and important ones. Note that the suggest numbers of SIC-95, SIC-99,[6] BIC, UAED and ENV are in line with this exhaustive analysis. ENV seems to provide a good compromise between SIC-95 and, on the opposite side, SIC-99 and BIC. The result provides by ENV is quite close to the UAED so that $k_e = 11$ with an high degree of reliability of $R_D = 0.86$.
Note that, in this experiment, the reliability measures provided by SIC shown that SIC is more "optimistic" in suggestion a more parsimonious model: for instance, the choice $k_e = 7$ is fostered by SIC with 95% of reliability, whereas CI = 84% and $R_D = 55\%$. Specially, the R_D index warns that with $k_e = 7$ we are missing 45% of the effective number of variables (the ENV index is 12.74).

5.3 Variable selection in a biomedical classification problem with real data

In [47], a feature selection analysis has been performed in order to find the most important variable for predicting patients at risk of developing nonalcoholic fatty liver disease among 35 possible features. The authors have collected data from 1525 patients who attended the Cardiovascular Risk Unit of Mostoles University Hospital (Madrid, Spain) from 2005 to 2021, and use a random forest (RF) technique as classifier and to yield a ranking of the components of input vectors (of dimension 35). They found that 4 features were the most relevant according to the ranking and considering the medical experts' opinions: (a) insulin resistance, (b) ferritin, (c) serum levels of insulin, and (d) triglycerides.
Here, we consider $V(k) = 1 - \text{accuracy-in-class}(k)$ as a figure of merit (where we set $V(0) = 0.5$,

[6] The numbers 95 and 99 are associated to the sum of some normalized weights (0.95 and 0.99) built by SIC which delivers a degree of *reliability* in the decision, as a confidence measure [23]: closer to 100, the decision is more *safety*.

that represents a completely random binary classification). The curve is obtained after ranking the 35 features [47]-[23, Section 5.5]. Note that with this choice of $V(k)$ we cannot apply BIC, AIC and other standard information criteria, as shown in Table 1. However, UAED [26], SIC [23] and ENV can be applied. The results are given in the Table below:

Scheme	UAED	SIC-90	SIC-95	SIC-99	**ENV**
k_e	3	2	3	9	**4**
CI	0.91	0.88	0.91	0.98	0.92
R_D	0.84	0.56	0.84	1	1

ENV gives a value of 3.5851, then suggests $k^* = 4$ as the experts in [47]. UAED and SIC also provide results in line with this solution. In this experiment, in terms of reliability, the values of the reliability measures provided by SIC and the measures proposed here are very close: for the decision $k_e = 3$ SIC provides safety measure of 0.95, and CI = 0.91, $R_D = 0.84$ that are in the same line (high values similar to 0.95). The same observation can be done for $k_e = 2, 4$ and 9 (with the exception of the R_D value - 0.56 - for $k_e = 2$, that alerts of a risky decision missing almost the 50% of the effective number of variables).

6 Conclusions

An index of effective number of variables has been proposed which has been inspired by the concept of maximum AUC) in ROC curves and the Gini index. The introduced ENV index removes a dependence that we can find in the elbow detectors designed in the literature, i.e., the dependence on the maximum number of components K analyzed. We also introduce different measures of uncertainty and reliability of the proposed solution (related to the ENV index). These novel reliability measures can be employed also jointly with the use different information criteria such as the well-known AIC and BIC (where they can be applied, as we have shown in Section 5.2). Several comparisons with classical and recent schemes are provided in different experiments involving real datasets: we analyze two variable selection problem, one of them regarding a regression analysis and the other one involving a biomedical classification study. Related Matlab code is provided.

Acknowledgement

The work was partially supported by the Young Researchers R&D Project, ref. num. F861 (AUTO-BA-GRAPH) funded by Community of Madrid and Rey Juan Carlos University, and by Agencia Estatal de Investigación AEI (project SP-GRAPH, ref. num. PID2019-105032GB-I00).

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] 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.

[6] 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.

[7] 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.

[8] R. San Millán-Castillo, 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.

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

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

[11] 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.

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

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

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

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

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

[17] 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.

[18] 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.

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

[20] 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.

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

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

[23] 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.

[24] J. J. Dziak, D. L. Coffman, S. T. Lanza, R. Li, and L. S. Jermiin, "Sensitivity and specificity of information criteria," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 553–565, 03 2020.

[25] F. Llorente, L. Martino, E. Curbelo, J. Lopez-Santiago, and D. Delgado, "On the safe use of prior densities for bayesian model selection," *WIREs Computational Statistics*, p. e1595, 2022.

[26] 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.

[27] 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.

[28] 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.

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

[30] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 3, pp. 267–276, 1953.

[31] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection - a review and recommendations for the practicing statistician," *Biometrical journal*, vol. 60, no. 3, pp. 431–449, 2018.

[32] 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.

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

[34] 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.

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

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

[37] L. Martino, V. Elvira, and F. Louzada, "Effective sample size for importance sampling based on discrepancy measures," *Signal Processing*, vol. 131, pp. 386–401, 2017.

[38] V. Elvira, L. Martino, and C. P. Robert, "Rethinking the Effective Sample Size," *International Statistical Review*, vol. 90, no. 3, pp. 525–550, 2022.

[39] 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.

[40] D. Watson, J. O' Hara, N. Tax, R. Mudd, and I. Guy, "Explaining predictive uncertainty with information theoretic shapley values," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, 2023, pp. 7330–7350.

[41] K. Aas, M. Jullum, and A. Loland, "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values," *Artificial Intelligence*, vol. 298, p. 103502, 2021.

[42] Wikipedia, "Bayesian information criterion," 2024, see the section 'Gaussian special case'. [Online]. Available: https://en.wikipedia.org/wiki/Bayesian_information_criterion

[43] 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.

[44] J. Fan, M. Thorogood, and P. Pasquier, "Emo-soundscapes: A dataset for soundscape emotion recognition," in *2017 Seventh international conference on affective computing and intelligent interaction (ACII)*, 2017, pp. 196–201.

[45] M. Efroymson, "Multiple regression analysis," *Mathematical methods for digital computers*, pp. 191–203, 1960.

[46] R. R. Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, pp. 1–49, 1976.

[47] 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.